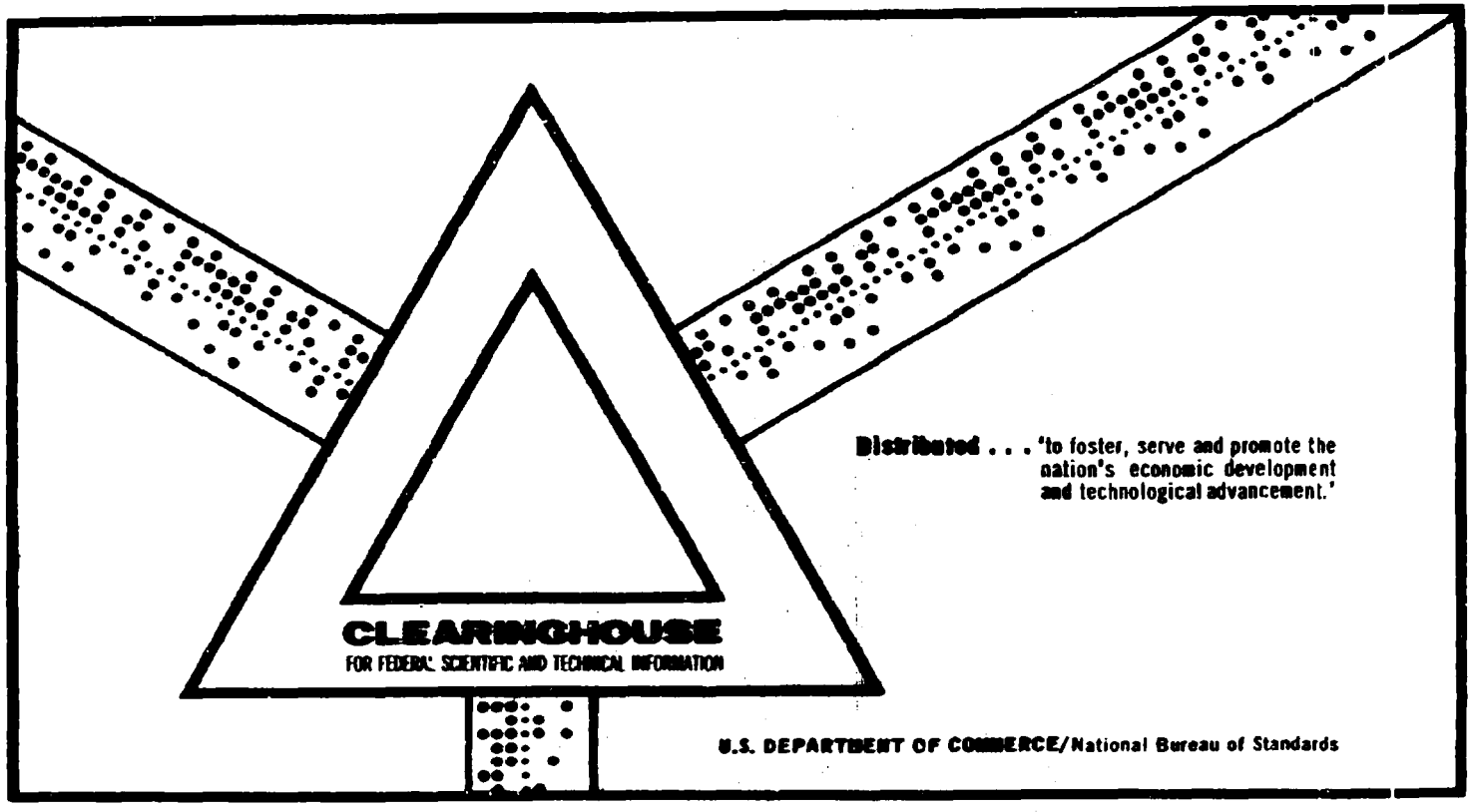
ISODATA, A NOVEL METHOD OF DATA ANALYSIS AND PATTERN CLASSIFICATION

Geoffrey H. Ball, et al

Stanford Research Institute
Menlo Park, California

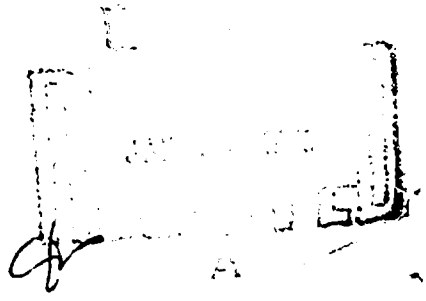
April 1965



ISODATA, A NOVEL METHOD OF DATA ANALYSIS AND PATTERN CLASSIFICATION

By: GEOFFREY H. BALL AND DAVID J. HALL

Technical
Report
April 1965



STANFORD RESEARCH INSTITUTE

MENLO PARK, CALIFORNIA

Prepared for
Information Sciences Branch
Office of Naval Research
Contract Nonr 4918(00)
SRI Project 5533

ISODATA, A NOVEL METHOD OF DATA ANALYSIS AND PATTERN CLASSIFICATION

By: GEOFFREY H. BALL AND DAVID J. HALL

Technical
Report
April 1965

Prepared for
Information Sciences Branch
Office of Naval Research
Contract Nonr 4918(00)
SRI Project 5533

CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (continued)

ABSTRACT

ISODATA, a novel method of data analysis and pattern classification, is described in verbal and pictorial terms, in terms of a two-dimensional example, and by giving the mathematical calculations that the method uses. The technique clusters many-variable data around points in the data's original high-dimensional space and by doing so provides a useful description of the data. A brief summary of results from analyzing alphanumeric, gaussian, sociological and meteorological data is given.

In the appendix, generalizations of the existing technique to clustering around lines and planes are discussed and a tentative algorithm for clustering around lines is given.

ISODATA, A NOVEL METHOD OF DATA ANALYSIS AND PATTERN CLASSIFICATION

by

Geoffrey H. Ball and David J. Hall
Stanford Research Institute

I INTRODUCTION

In this paper we discuss a technique for dealing with problems in which the data are inherently described in many dimensions, where each dimension corresponds to a variable of the problem. Such problems are very commonplace,* and many which are customarily described in terms of a small (3 to 10) number of dimensions are in fact of a much larger dimensionality, but have been simplified in order to allow manipulation (and description) of the data. Such collapsing of the problem is often useful and suitable, indicating the preponderant importance of certain of the parameters. However, there remains a class of problems for which such collapsing destroys significant interrelations between the parameters that would give the data meaning.

II PATTERN RECOGNITION PREPROCESSING AND CLASSIFICATION

The area of research labelled "pattern recognition" consists primarily of efforts to develop techniques capable of dealing with problems of inherently high dimension.

Many aspects of the pattern recognition problem are, in fact, data analysis called by a different name. Realizing this, we feel free to discuss ISODATA in the context of pattern recognition (which is our background), although others might prefer "automatic data analysis" as the label for our work.

One statistician, John W. Tukey, has stated that, in his mind, data analysis includes, among other things: "procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise and more accurate, and all the machinery and results of (mathematical)

*Examples of such problems include automatic speech recognition, medical diagnosis, alpha-numeric character recognition, sociological questionnaire analysis, and weather prediction.

statistics which apply to analyzing data."[1]

Dr. Tukey goes on to say[2] that he considers data analysis as a science in the sense that it has:

a) intellectual content

b) organization into an understandable form

c) reliance upon the test of experience as the ultimate standard of validity.

As a science he feels that,

a) "Data analysis must seek for scope and usefulness rather than security.

b) Data analysis must be willing to err moderately often in order that inadequate evidence shall more often suggest the right answer.

c) Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proof of validity."

We feel that our work on ISODATA fits this description of data analysis. Since we have also been able to describe ISODATA in the terminology of pattern recognition, we feel justified in relating data analysis to pattern recognition. Our work on ISODATA-POINTS has concentrated primarily on developing the algorithm and demonstrating experimentally that it works on both real and artificial data. We are now engaged in studying the algorithm analytically and in comparing it to both other clustering techniques and to existing multivariate statistical methods.

A convenient (though not usually well-defined) division of the pattern recognition problems is:

1) Design and Evaluation of Transducers--here "transducers" are those parts of a pattern recognition system that transform a physical phenomenon into a set of electrical measurements or optical patterns that are in a form suitable for the preprocessor.

2) Design, (Automatic) Synthesis, and Evaluation of Preprocessing--here "preprocessing" is that part of the system that transforms the measurements from the transducers into multi-dimensional patterns. This transformation should enhance the differences between classes we want to discriminate. At the same time it should preserve within-class similarities.

3) Design, (Automatic) Synthesis, and Evaluation of the Categorizer--here the categorizer is that part that transforms the patterns from the preprocessing into labels (output codes) associated with each set or class of patterns.

In the design of the "preprocessing" to do automatic classification, a first sub-task is to analyze a representative set of patterns. Many present preprocessing evaluation techniques, even when used on a set of representative data, give little information as to how to modify the preprocessing technique in order to improve it.

In order to know whether we can improve the preprocessing we must be able to distinguish between:

1) poor performance of the preprocessor, and

2) inherently difficult data.

The proposed technique has demonstrated its ability to lay bare the structure of the data. ISODATA can be used to evaluate preprocessing by comparing the clustering before the preprocessing with the clustering after the preprocessing. This makes it possible to evaluate the preprocessor with respect to the inherent difficulty of the data. The clear picture of the data that the researcher obtains using this technique helps him to modify the preprocessor so that the resulting patterns are structured in a more suitable manner. We have found that understanding the structure of the data is also suggestive of new ways for transducing the physical phenomenon.

In the classification of patterns a primary problem is the obtaining of an economical description of the patterns. Due to the complex nature of patterns arising from the real world, a description not rigidly constrained by a number of a priori assumptions is desirable. The class of techniques we describe has the characteristic that its description of the data is determined primarily by the data and is self-adjusting in a way that makes this description economical.

In discussing this technique we concentrate on using it for the analysis of data. We see the technique as specifically applicable to:

1) The design of preprocessing (and also more conventional data analysis), by allowing the examination of the structure of multidimensional data in the original high-dimensional space; and

2) Classification of patterns, by finding the structure in each class of patterns and providing an economical description of the classes of patterns against which a pattern of unknown class can be compared and so assigned to a specific class.

III THE TECHNIQUE

ISODATA, as the derivation* of the name suggests, is a collection of iterative techniques. It does not attempt to summarize all of the finest nuances extractable from the data. Rather it focuses on central tendencies and the major structure of the data. As ISODATA is normally used, it is a compromise between attempting to store and analyze every detail and aspect of the data right from the start (if desirable, this can be done later in the analysis on limited portions of the data) and an approach that averages virtually everything together. (In fact, if it were desirable for some reason, either extreme is attainable by appropriate selection of certain process parameters that control ISODATA.)

It is not practical to compare all patterns with all other patterns for large numbers of patterns. Rather the procedure compares patterns with a set of clusters constructed from subsets of the patterns themselves, and groups patterns together on the basis of these comparisons. The comparisons are made by establishing a measure of distance in the measurement space. Patterns are grouped together if they lie closest to the same "description of a cluster."** The number of clusters used by the technique varies in a way that depends on the structure of the patterns in measurement space and on the ISODATA process parameters that the researcher controls.

When used on data for which categorization information is not available, ISODATA finds a good approximation to the natural structure of the data, rather than trying to impose an assumed structure on the data. By clustering only one class of patterns at a time, categorization information can be used in conjunction with ISODATA to structure the data for a specific pattern classification problem; a probability distribution of the data need not be known or even assumed to exist. The development of a computationally-simple method that could be implemented for patterns of more than 100 dimensions (e.g., optical patterns and complex waveforms) was an important factor guiding the development of this technique.

The simplest type of ISODATA is ISODATA-POINTS. This technique is described in the next section. In Appendix A we discuss ISODATA-LINES and ISODATA-PLANES, two generalizations of ISODATA-POINTS. There we give a tentative algorithm for ISODATA-LINES. Though neither of these generalizations has yet been programmed, we feel them to be relatively straightforward generalizations of ISODATA-POINTS.

---

*With apologies for adding another acronym to the growing list, we have coined ISODATA to represent Iterative Self-Organizing Data Analysis Techniques A. (The "A" was added to make ISODATA pronouncable.) The classically-oriented can derive it from ISO, meaning "the same" or "like," + Data.)

**Here we use "cluster" in a general way--allowing it to mean a set of patterns grouped around a point, a line, or a plane. Hence the "description" is the specification of the point, line, or plane around which the patterns are clustered.

IV DETAILED DESCRIPTION OF ISODATA-POINTS

In this section we describe ISODATA-POINTS from four points of view, each succeeding point of view being more precise than the last. The four points of view are:

(1) Verbal

(2) Pictorial

(3) Two-dimensional Illustrative Example

(4) Mathematical

We also give the results of a principal components analysis (a more conventional statistical analysis technique) on the same data so that the results of it can be compared with the type of results obtained using ISODATA-POINTS.

A. Verbal Description

ISODATA-POINTS is an iterative procedure for the sorting of a set of multi-dimensional (multi-variable) patterns into subsets of patterns. An average pattern is used to represent each subset of patterns, and the iterative process, by changing the composition of these subsets, creates new average patterns. These new average patterns define new subsets each of which has reduced variation about the average pattern. The process also combines average patterns that are so similar that their being separate fails to provide a significant amount of additional information about the structure of the patterns.

B. Pictorial Flow Diagram

We show a pictorial flow diagram of ISODATA-POINTS in Fig. 1. In line with our considering ISODATA as a procedure for sorting patterns we show the patterns being fed into a sorter, one at a time, from a "pattern hopper." The patterns are sorted into subsets on the basis of distance from a set of cluster points--each pattern going into that subset associated with the cluster point to which it, the pattern, is closest. The cluster points themselves are obtained as an output of the previous iteration. The set of cluster points for the first iteration must be provided by the researcher.*

*The selection can be arbitrary, since the results of clustering have been found experimentally to be nearly independent of the choice of the initial cluster points. Usually, however, a wise choice reduces significantly the number of iterations needed for satisfactory clustering. We have found it best to use a subset of the patterns randomly selected from the training patterns as the initial cluster points.

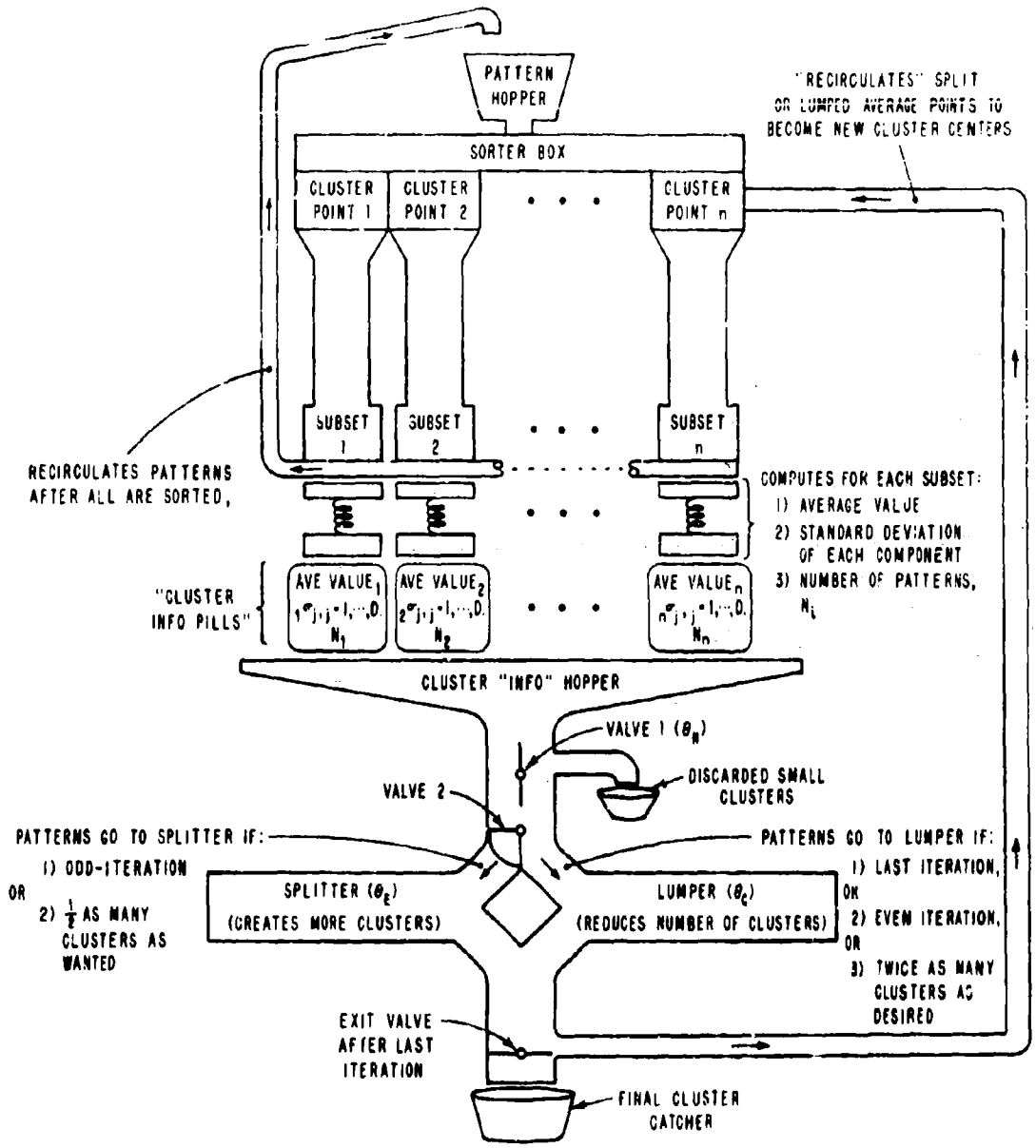


FIG. 1 A PICTORIAL DESCRIPTION OF ISODATA-POINTS

After all patterns have been sorted, the average of each of the subsets of patterns is computed, and the sample standard deviations in each dimension of each subset are determined. The average pattern vector, the standard deviation in each dimension for each subset, and the number of patterns in each subset are then passed on into the "Cluster Information Hopper."

Those small clusters (with fewer than θ_N elements) are discarded at "Valve 1." The positioning of "Valve 2" is determined by the number of the iteration and by the total number of clusters, as indicated on the diagram.

The criteria and method of splitting and lumping of clusters are given in detail in the next two sections. Splitting takes place if the standard deviation in any dimension is greater than θ_S and also if both (1) the cluster has enough members to split and (2) has high average distance between its mean and the patterns in its subset.

Lumping occurs between, at most, the L pairs of means that are less than θ_C apart. (The process parameters θ_S, θ_C, θ_N, and L are all supplied by the researcher.)

After each lumping or splitting, the modified set of average points is used as the set of cluster points for the next iteration and placed in the "Sorter Box." The program ends when the number of iterations performed equals the number specified by the researcher. At this stage, the cluster points should adequately "fit" the data.

C. Two-Dimension Illustrative Example

In order to illustrate the details of ISODATA-POINTS we have contrived the set of two-dimensional patterns shown in Fig. 2. The two dimensions are height and weight. The patterns (the points shown in Fig. 2) are intended to represent the height and weight of typical professional athletes.

Given these points to cluster, the ISODATA-POINTS technique proceeds in a manner that we illustrate in Figs. 2-19. Each figure illustrates a major step in the computer program. (The actual figures are placed after the explanatory text for all of the figures and can be folded out.)

The clustering shown in this example was found by running the existing ISODATA-POINTS computer program. One particularly interesting aspect of this run is the way in which the technique found and isolated a number of points lying virtually alone. This offers one approach to treating "wild shots" in the analysis of data, since they are singled out for further study or for discarding.

Figure	Step
2	Selection of the Pattern Set

Note that three distinct clusters are labelled "Rugby" and two are labelled "Basketball." Therefore, for this problem a simple average, for example of all basketball players, will not describe satisfactorily a representative basketball player.* In other words, the classes are composed of several subclasses, i.e., they are "multi-modal."

Obviously this particular data set can be analyzed with pencil and paper. A computer is not necessary because the data points are described by only two characteristics of the athletes, i.e., their weight and height. If, however, these data were described in many more dimensions, e.g., more than 3 or 4, then it becomes nearly impossible to display them satisfactorily in their raw form even in many sets of two-dimensional representations.

We will show in Section VI how to obtain for these cases a two-dimensional plot that usefully describes the data.

One important goal of this ISODATA analysis is a comprehensible and useful description of the data. In order to obtain this description we seek to divide the data points into relatively homogeneous subsets, each subset of which can be adequately described by its average point. The following example seeks to describe how relatively homogeneous subsets of data can be obtained.

3	Selection of Initial Cluster Points

Note that one small region has two initial cluster points and another small region has three. These initial trial cluster points were selected to show that if by arbitrary selection a bad choice of initial cluster points is made, that even then the final cluster points will be good ones.

4	Partitioning of the Pattern Space as Defined Implicitly by the Cluster Points

Note that the boundaries are the perpendicular bisectors of the lines joining pairs of cluster points. Since we are seeking minimum distance of a pattern over all cluster points, the boundaries are meaningless except where they are between the two closest

* Or to say this in another way, a man with his head in an oven and his feet on a cake of ice can hardly be adequately described as being "warm on the average."

Figure	Step
	cluster points. Hence the piecewise-linear nature of the boundary.
5	Sorting of the Patterns for Iteration 1 using the Initial Partition Note that the patterns are assigned to only one subset, and that all subsets are contained in convex volumes of pattern space. Note also that the initial partition is not a good one. Subsets having fewer than θ_N elements would be discarded at this point. (θ_N is a researcher-supplied process parameter.) It may be helpful in this particular example to think of the data points as representing men standing in a large field. The men are positioned in the field in accordance with their weight and their height. The partitions that divide the data points into subsets can be thought of as "fences" dividing the men into groups. The cluster points can be considered as "group leaders" to whom the men owe temporary allegiance, i.e., a man owes his allegiance to the closest group leader. As we shall see, in the ISODATA process group leaders come and go (it was ever thus).
6	Finding the Average Point of Each Subset After the first iteration the ISODATA average points become cluster points.
7	Splitting of the Average Points takes place when (1) the maximum standard deviation exceeds θ_s (a researcher-supplied ISODATA process parameter) and either 2 or 3 is true. (2) the number of patterns in a subset exceeds $(2\theta_N + 2)$ and (when the average distance of patterns in subset i from the average point of subset i) exceeds $\overline{AD}$, the average distance of a pattern from its closest average point. NROWS is the number of clusters

* A volume is convex if the straight line connecting any two points in the volume lies entirely within the volume.

Figure	Step
7 (Cont.d)	More precisely,

$$\overline{AD} = \frac{1}{N} \sum_{i=1}^{NROWS} (AVEDST_i) \times (N_i)$$

and

$$AVEDST_i = \frac{1}{N_i} \sum_{j=1}^{N_i} (\text{Distance of pattern } j \text{ from the mean vector of subset } i)$$

(For all patterns in subset i)

(3) The number of clusters is less than or equal to one-half the number of clusters that the researcher has specified as being desired.

The Splitting occurs in the following fashion:

If the conditions of splitting are satisfied then the first new average points (for example, the right hand or the upper average points of Fig. 7) are created by <u>adding</u> 1 to that component of the original average point having the largest standard deviation. The second "average points" (for example, the left hand or the lower average points of Fig. 7) are created by <u>subtracting</u> 1 from that component of the original average point having largest standard deviation.

---

* The actual amount added is arbitrary (here it is +1) so long as it is sufficient to provide detectable differences in the distance of a pattern from the two cluster centers and is not so large as to change other boundaries appreciably.

Figure	Step
8.	The Partition for Iteration 2. Note that the boundaries between the pairs of cluster points split from a single point are perpendicular* to the direction having maximum standard deviation.
9.	Sorting of the Patterns for Iteration 2.
10.	Finding the Average Points of Iteration 2. Note the effect of the "outlier" (shown with an arrow) on the average point for the uppermost cluster.
11.	The Average Points found in Iteration 3. In iteration 2 the average points were again split (since the number of subsets was less than one-half the number of subsets desired). Note that the "outlier" of Fig. 10 has been made a cluster by itself.
12.	The Average Points of Iteration 3 are Split in the Manner Described under Fig. 7 above.
13.	The Sorting of the Patterns in Iteration 4.
14.	The Finding of Average Points for Iteration 4.
15.	The Lumping Together of Close Average Points. In this iteration the criteria for lumping (an even iteration and the existence of more than one-half the number of subsets desired) are satisfied. This figure illustrates the lumping together of all pairs of average points that are less than a distance of θ_C apart. (θ_C is a researcher-supplied ISODATA process parameter). Note that only pairs** of average points are lumped together. Note also that the lumped average point obtained is the average of the two average points and is obtained by weighting each average point by the number of patterns in its subset. This makes the lumped average point the true average point of the combined subset.

*Splitting of average points in several dimensions (into more than two new "average points") was once considered for use in the algorithm. We found that this becomes hazardous unless the covariance matrix is calculated, and this calculation is undesirable.

**Lumping triples was considered for the algorithm, and discarded, since it appeared to change the partition too radically for the iterative procedure to satisfactorily "converge."

Figure	Step
16.	The Partition for Iteration 5.
17.	The Average Points for the Subsets of Iteration 5.
18.	The Average Points for Iteration 6.

In the previous iteration, six average points were split. In this iteration four pairs of average points will be lumped together. These four pairs of points are indicated by being circled.

19. The Final Average Points After Several Iterations.

No splitting or lumping is allowed in this final iteration, which is principally for consolidation. We feel that these 19 average points do quite adequately describe this set of 562 data points with a 30:1 reduction of the number of points. Naturally, no description of the original data points which provides a similar amount of reduction can be as accurate as the original data itself. However, a reduced data description is often more useful than the more accurate but much more voluminous version.

Note the way the "wild shots" or "outliers" have been found and isolated by assigning them to clusters of their own. These wild shots can now be examined for their importance--either as a rare occurrence well worth noting, or as an equipment malfunction.

We terminated the iterative procedure after iteration 7, because the clustering obtained seemed quite adequate. If it had been necessary, we could have gone on by continuing the lumping and splitting, starting at the end of iteration 6. Our experience has shown that about six iterations are adequate for many problems* --adequate in the sense that the number of clusters is stable and the subsets relatively homogeneous.

We can, by increasing Θ_C and increasing Θ_E reduce the number of clusters we obtain. Decreasing them both would increase the number of clusters.

The number by each average point gives the number of patterns in that cluster.

*We are presently seeking better criteria for terminating the iterations.

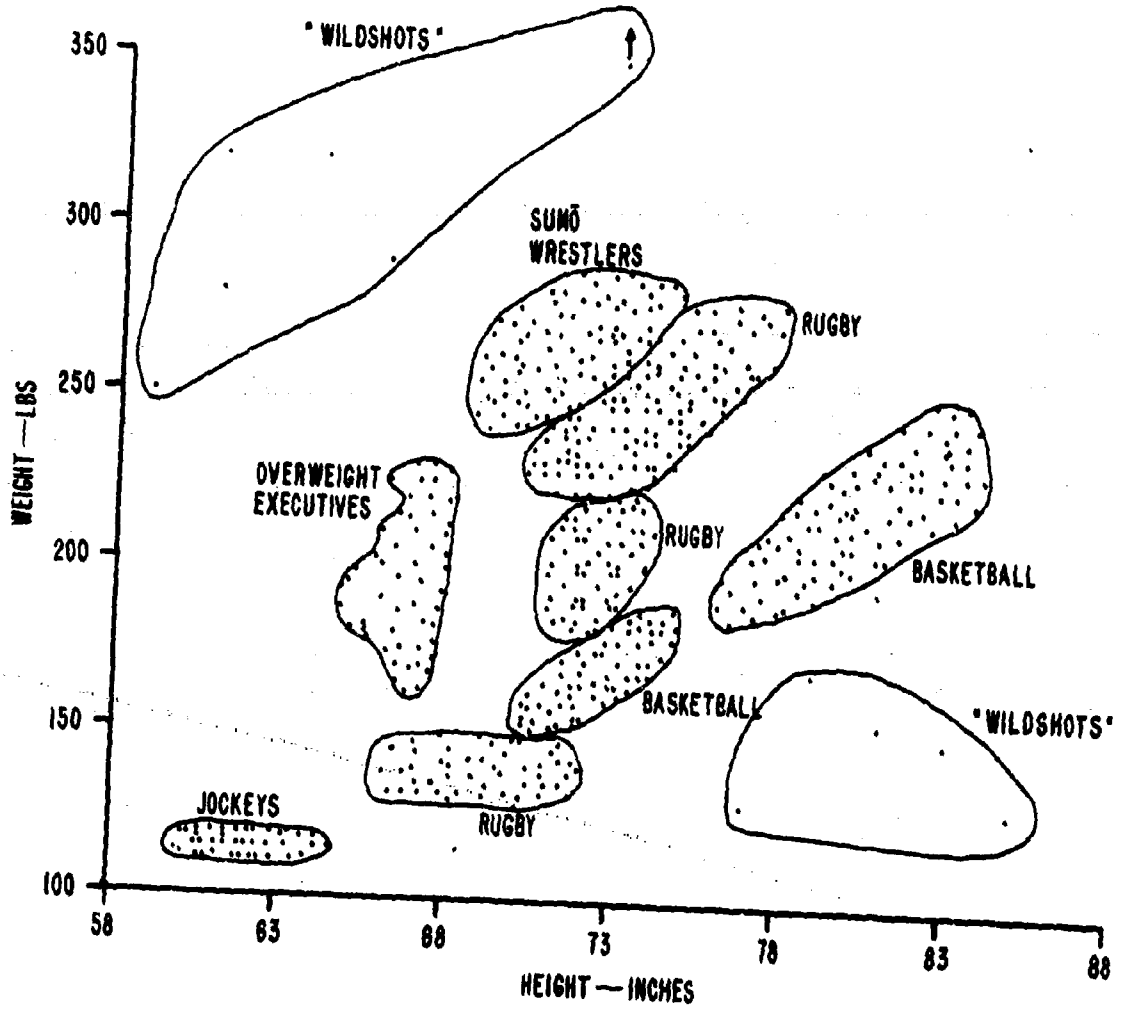


FIG. 2 SELECTION OF THE PATTERN SET

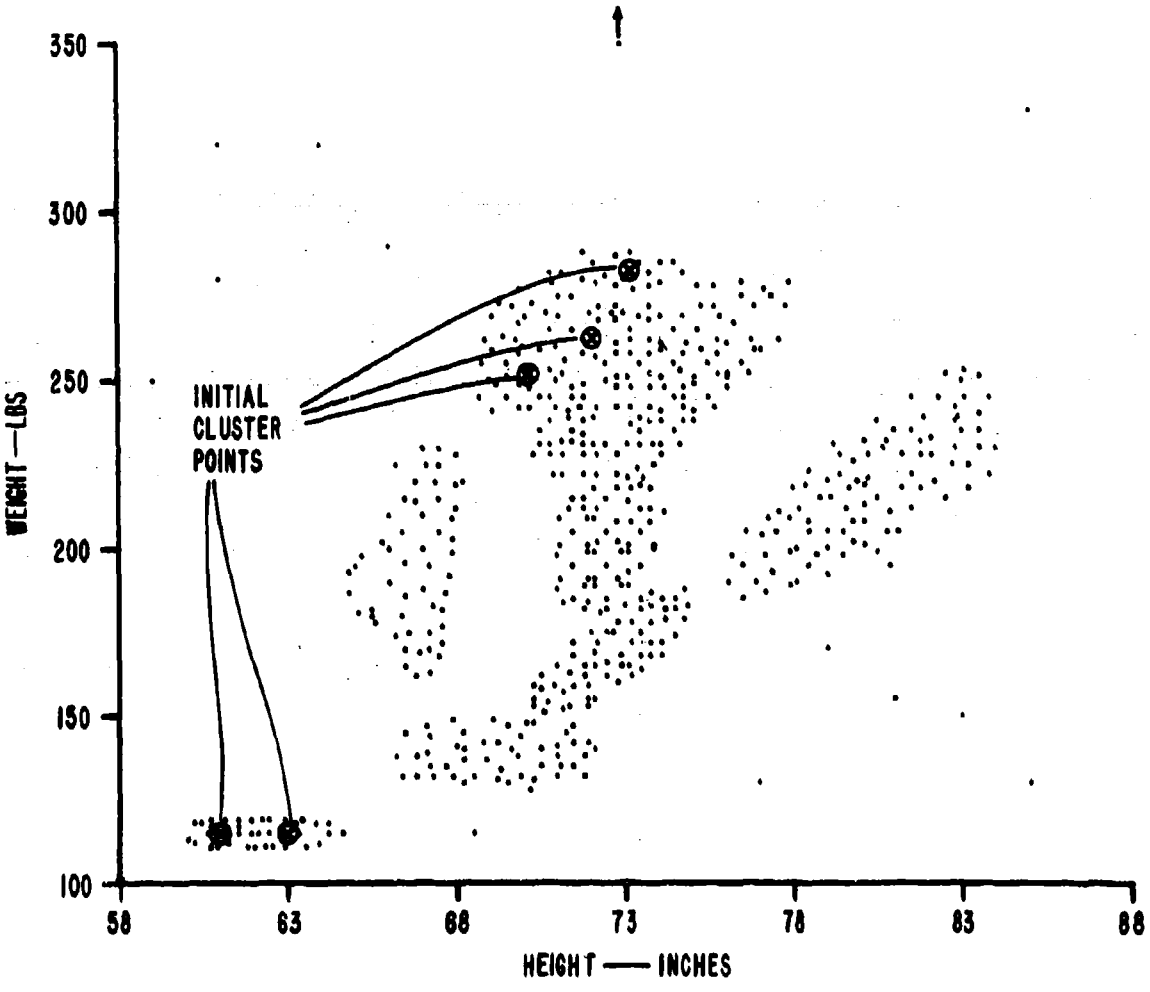


FIG. 3 SELECTION OF INITIAL CLUSTER POINTS

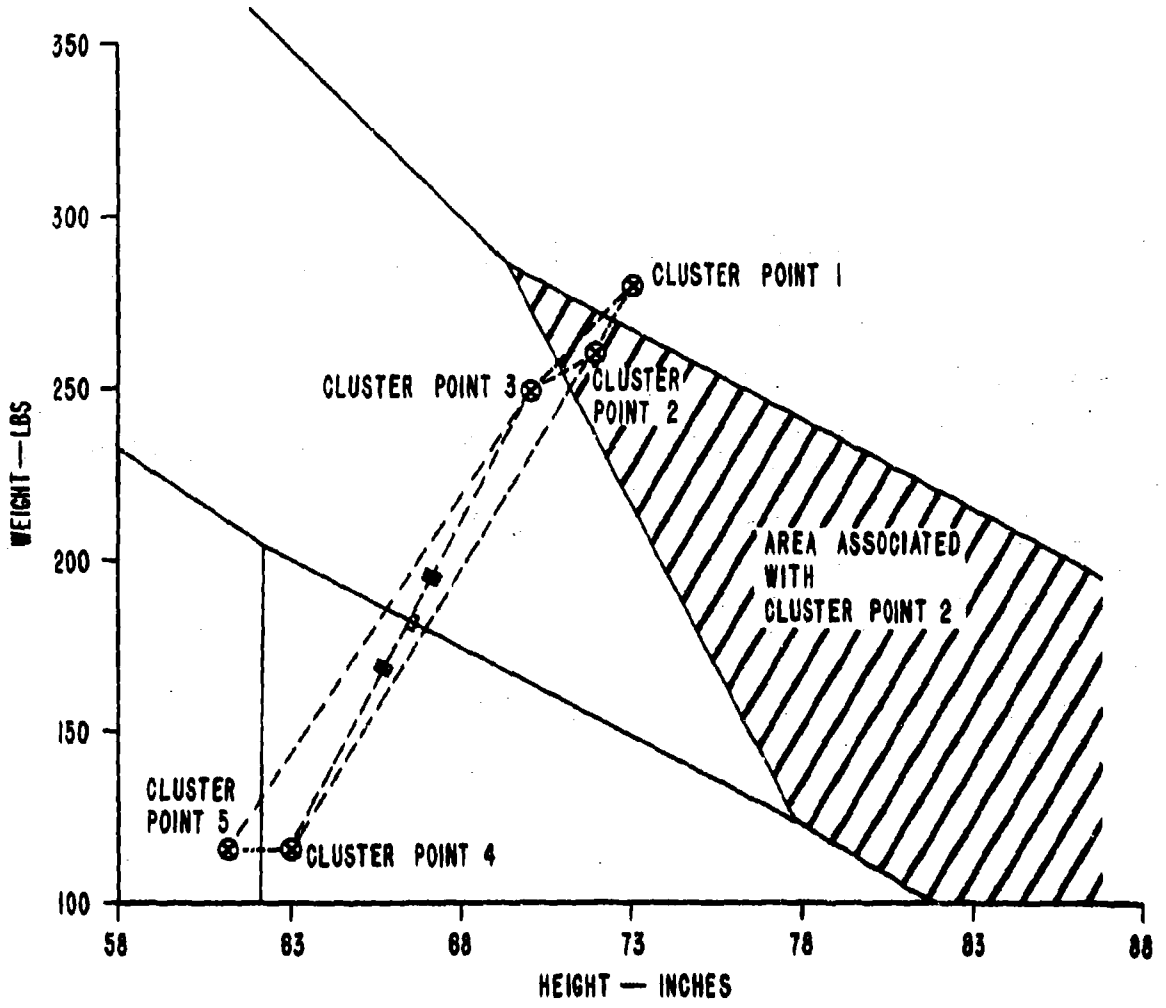


FIG. 4 PARTITIONING OF THE PATTERN SPACE AS DEFINED IMPLICITLY BY THE CLUSTER POINTS

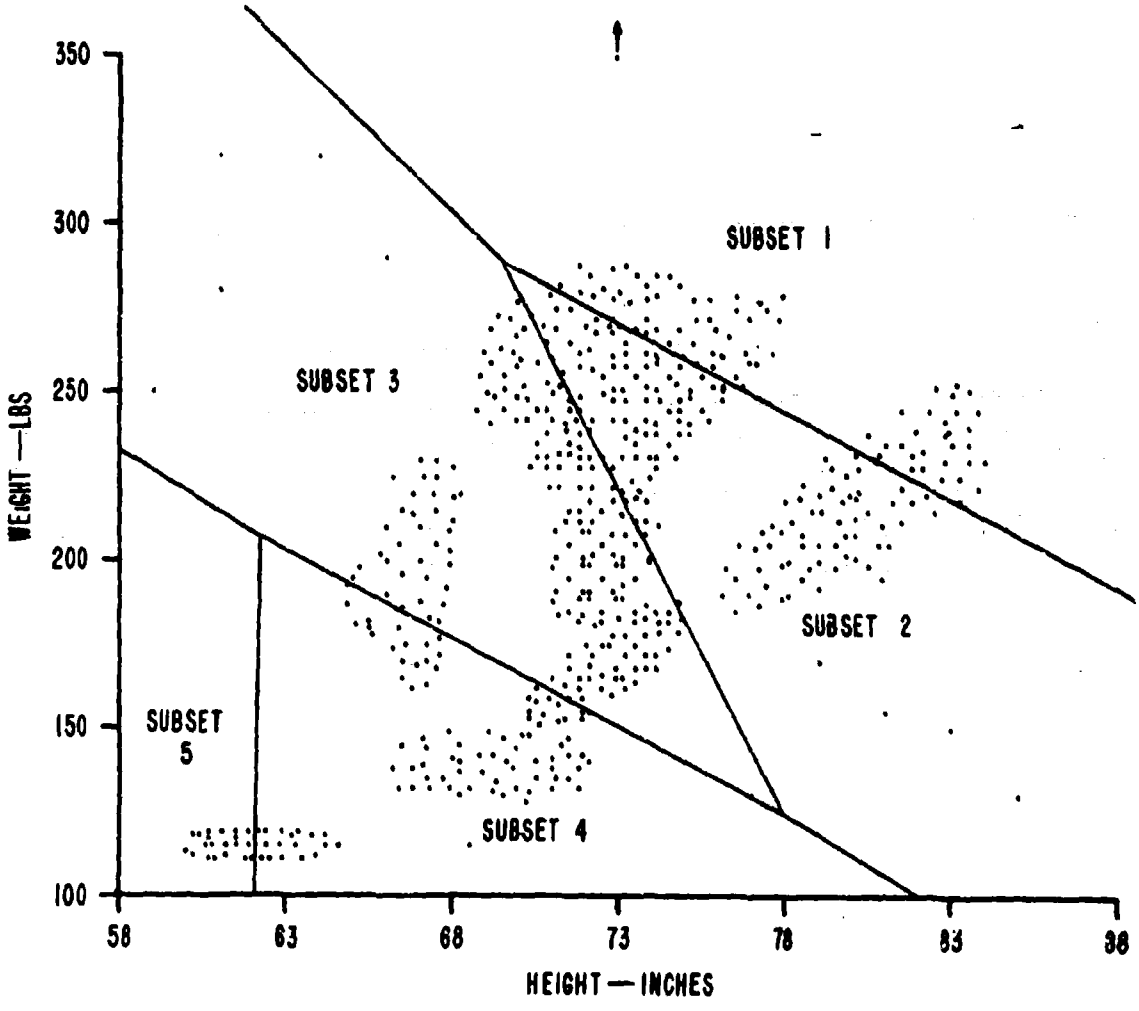


FIG. 5 SORTING OF THE PATTERNS FOR ITERATION 1 USING THE INITIAL PARTITION

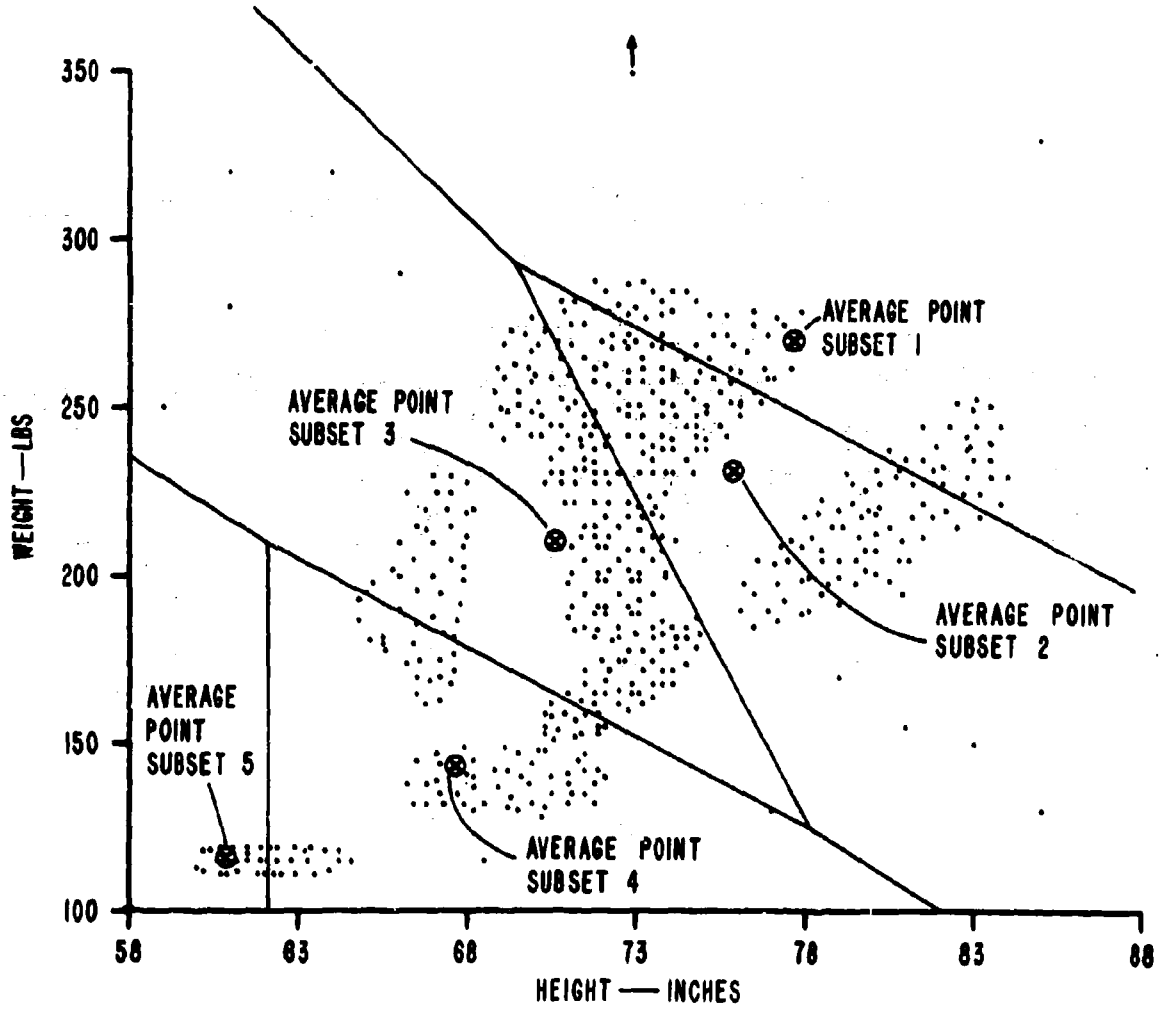


FIG. 6 FINDING THE AVERAGE POINT OF EACH SUBSET

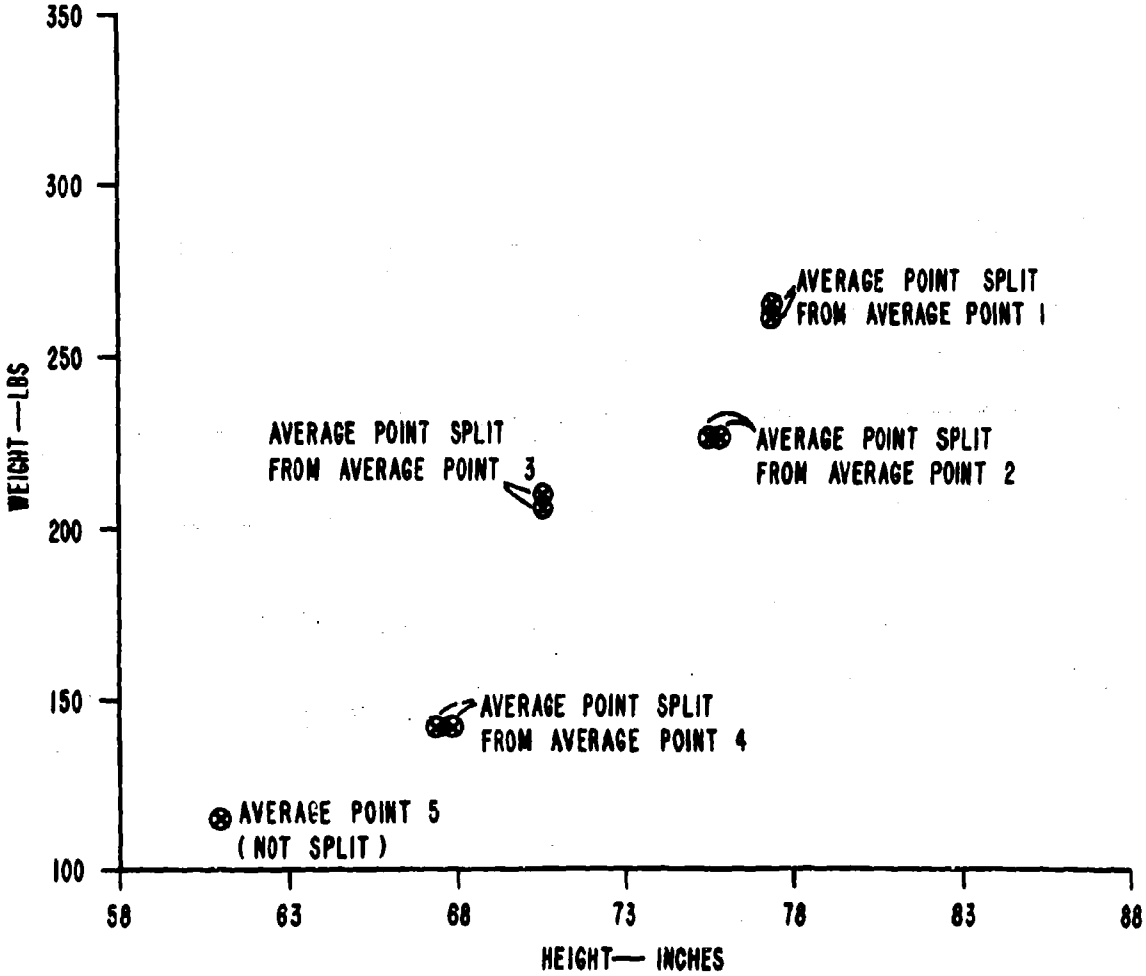


FIG. 7 SPLITTING OF THE AVERAGE POINTS

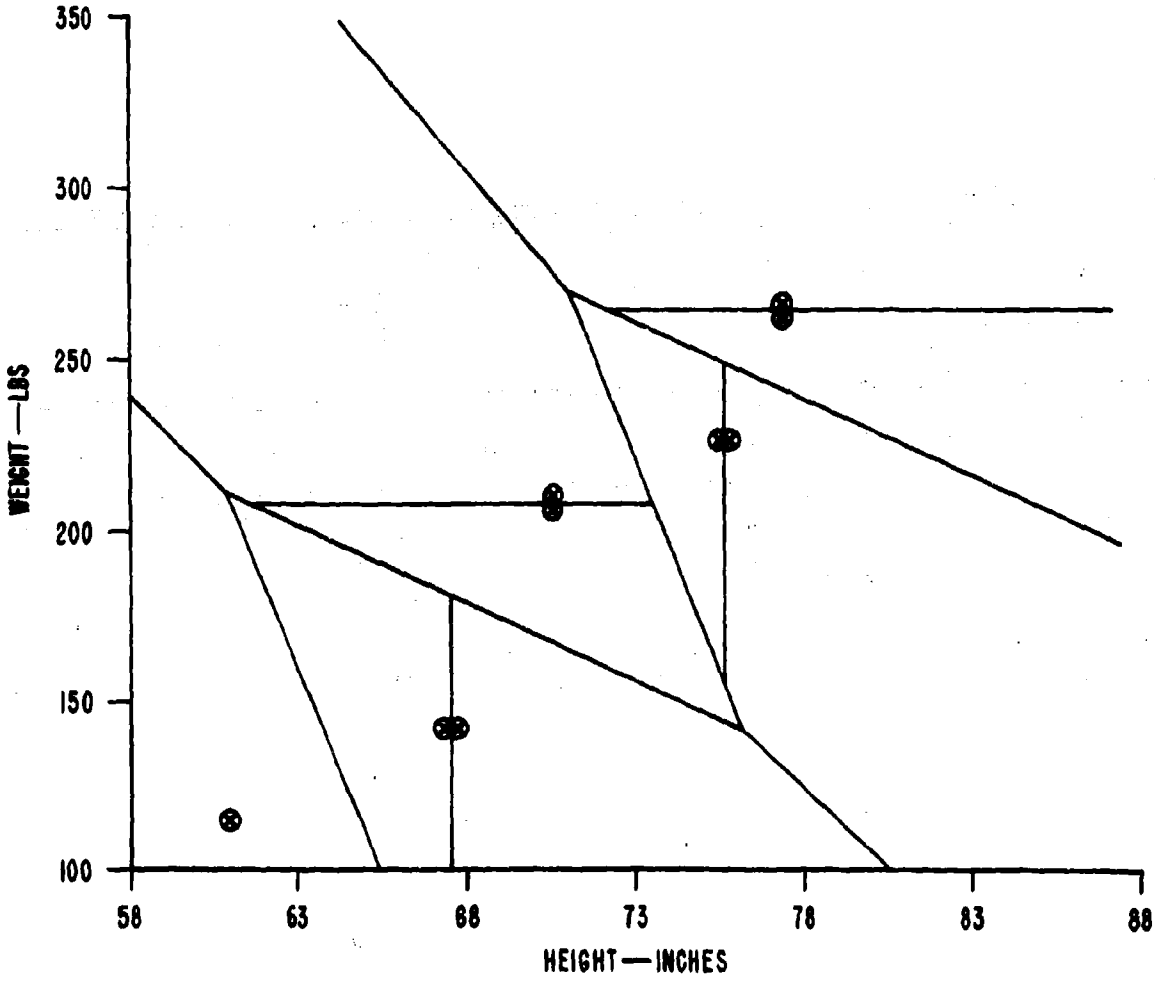


FIG. 8 THE PARTITION FOR ITERATION 2

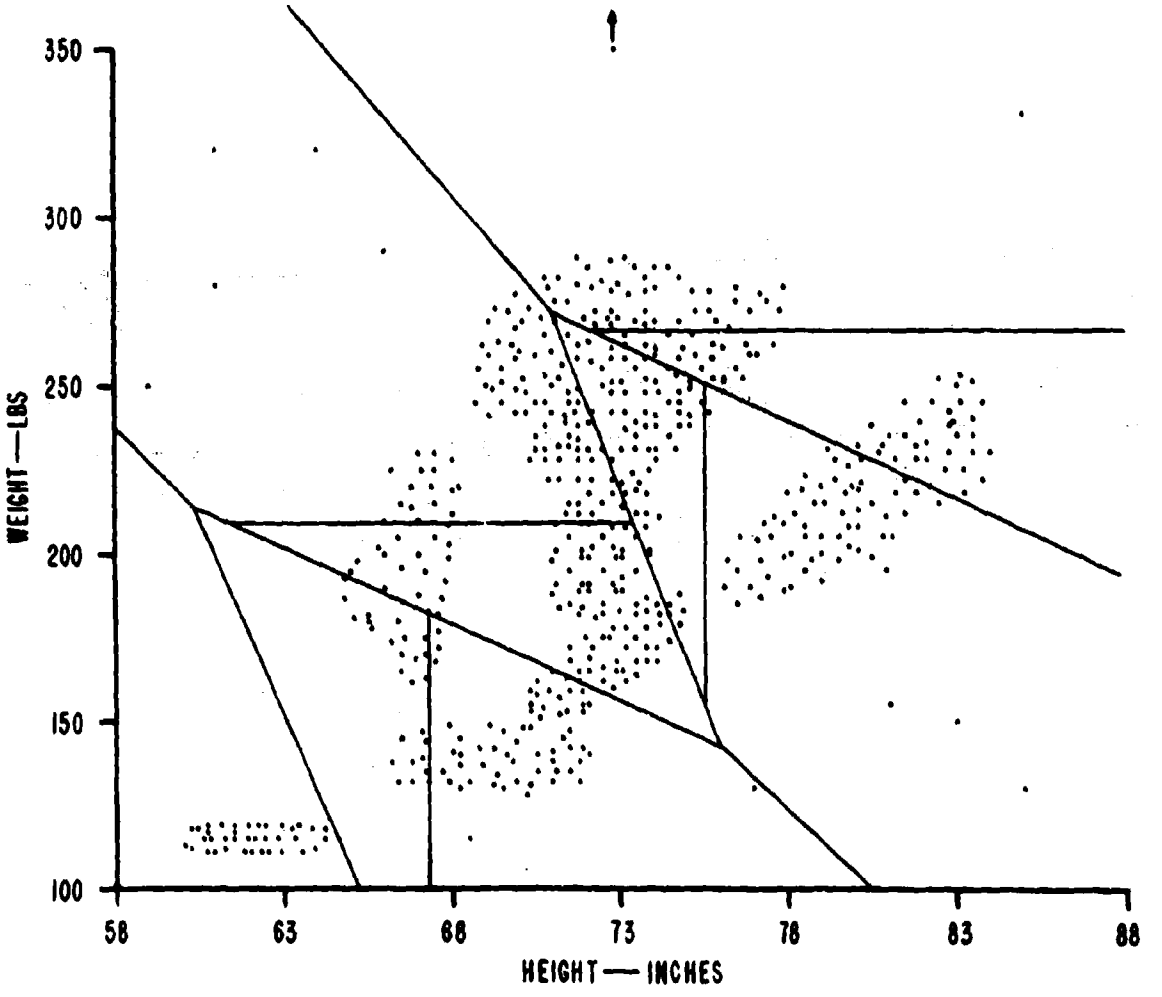


FIG. 9 SORTING OF THE PATTERNS FOR ITERATION 2

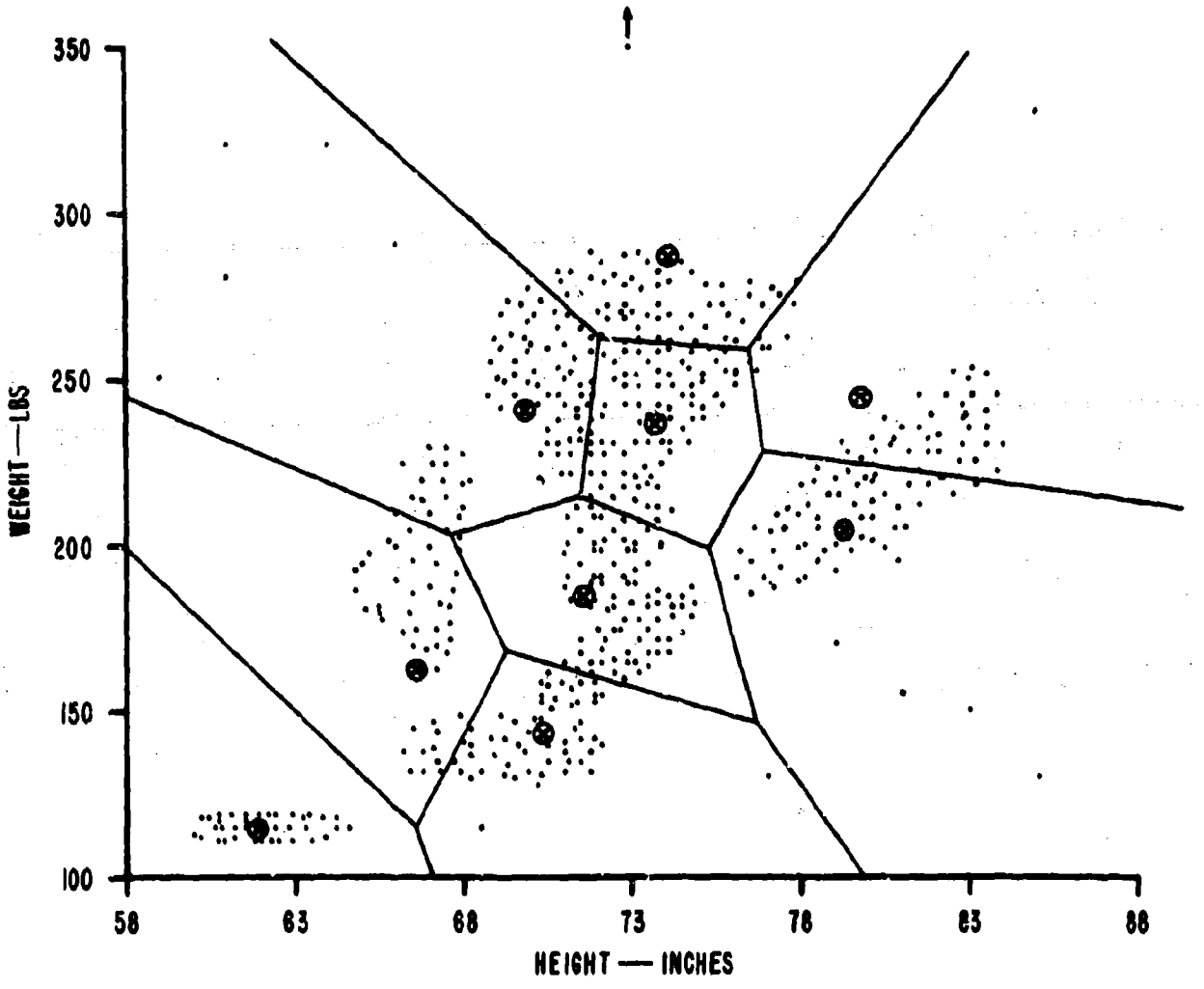


FIG. 10 FINDING THE AVERAGE POINTS OF ITERATION 2

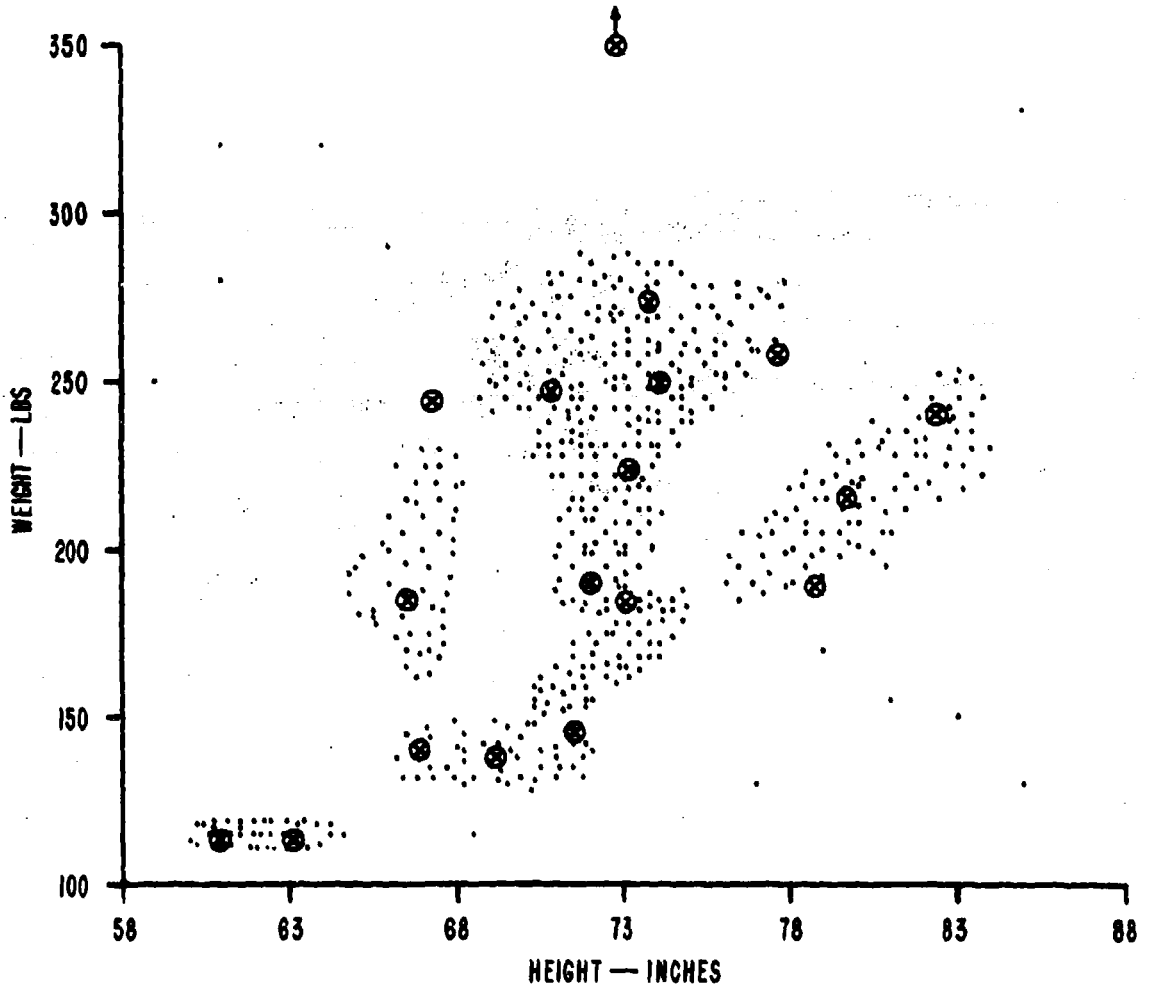


FIG. 11 THE AVERAGE POINTS FOUND IN ITERATION 3

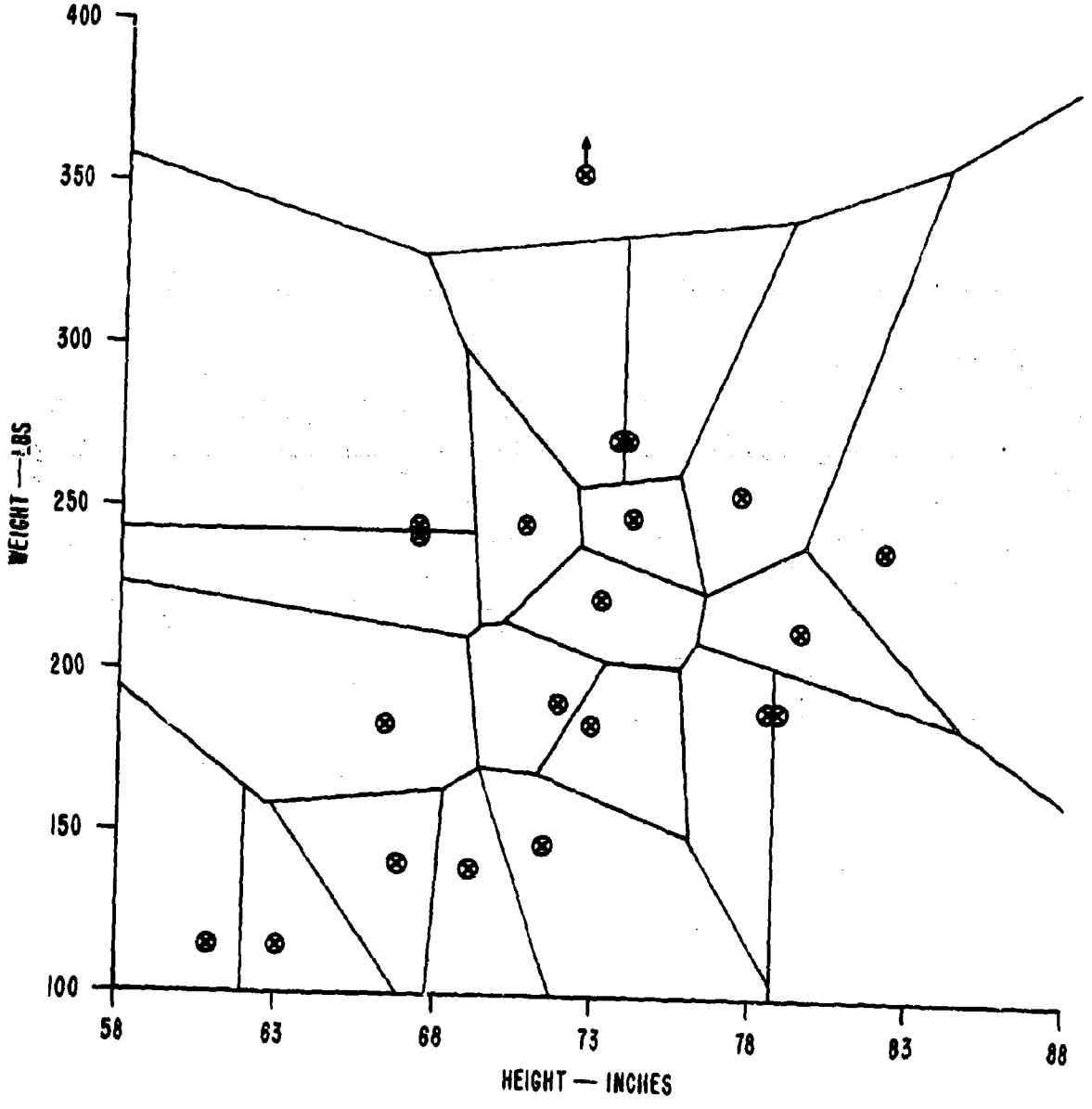


FIG. 12 THE AVERAGE POINTS OF ITERATION 3 ADR SPLIT IN THE MANNER DESCRIBED UNDER FIG. 7 ABOVE

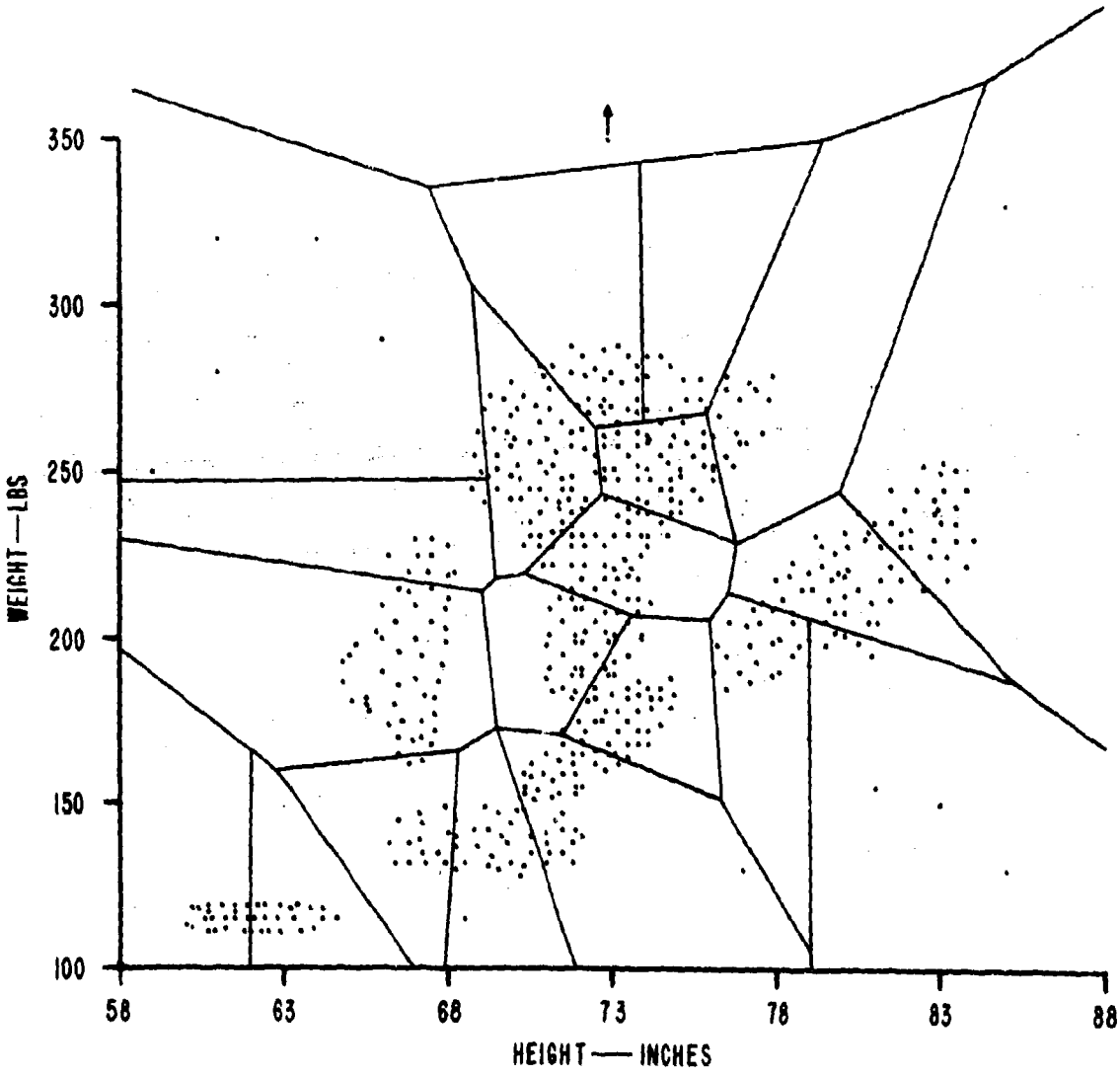


FIG. 13 THE SORTING OF THE PATTERNS IN ITERATION 4

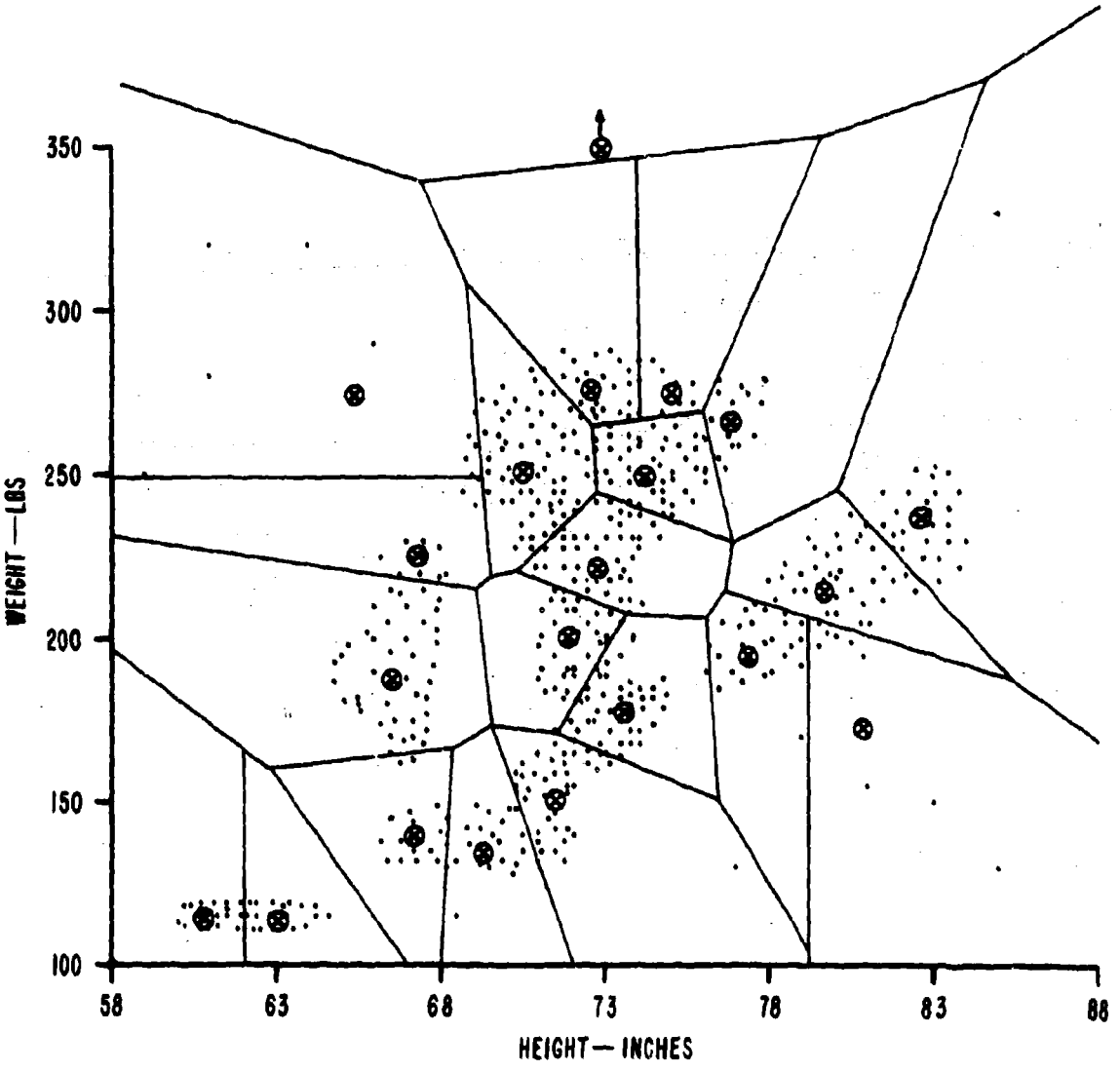


FIG. 14 THE FINDING OF AVERAGE POINTS FOR ITERATION 4

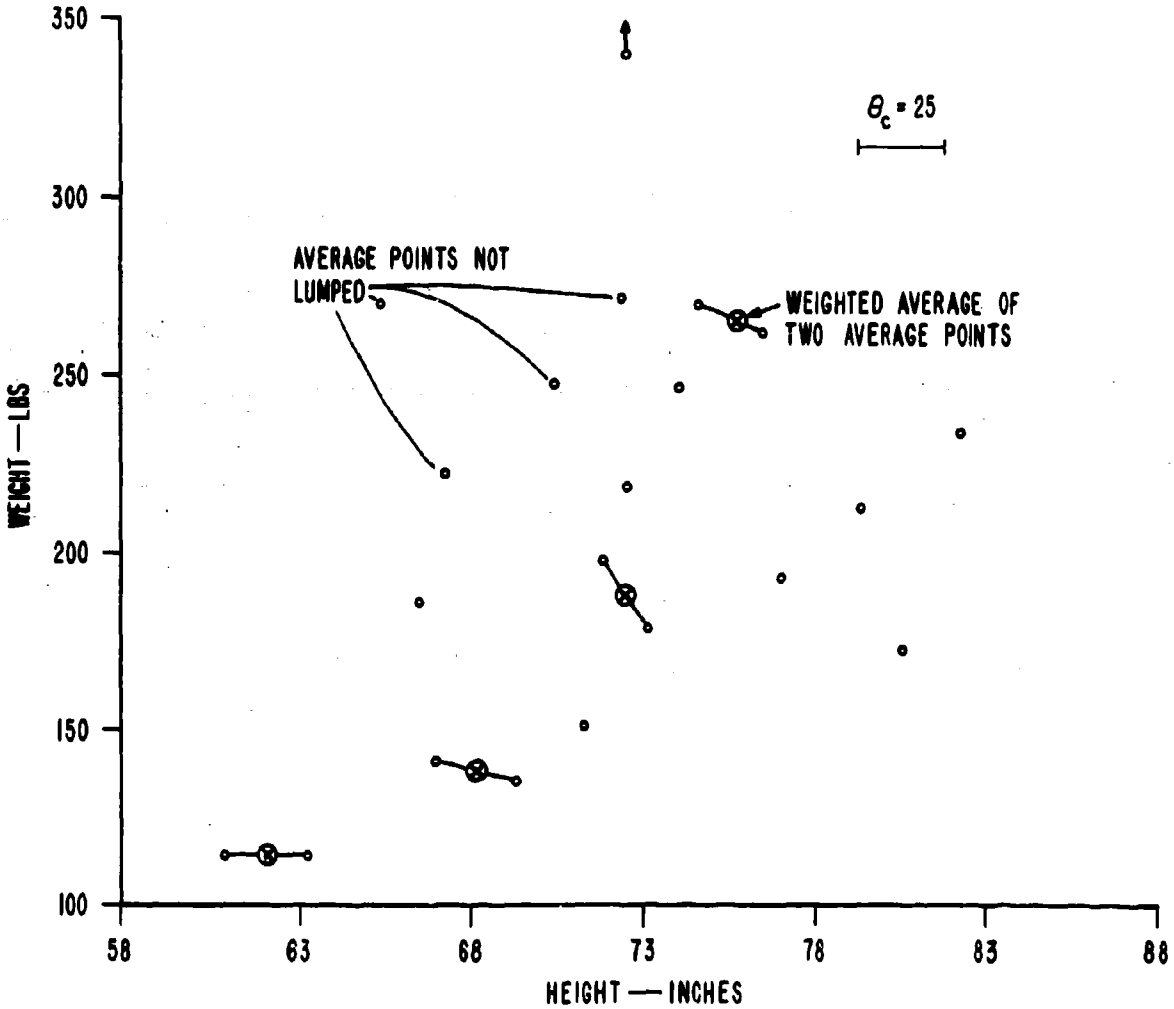


FIG. 15 THE LUMPING TOGETHER OF CLOSE AVERAGE POINTS

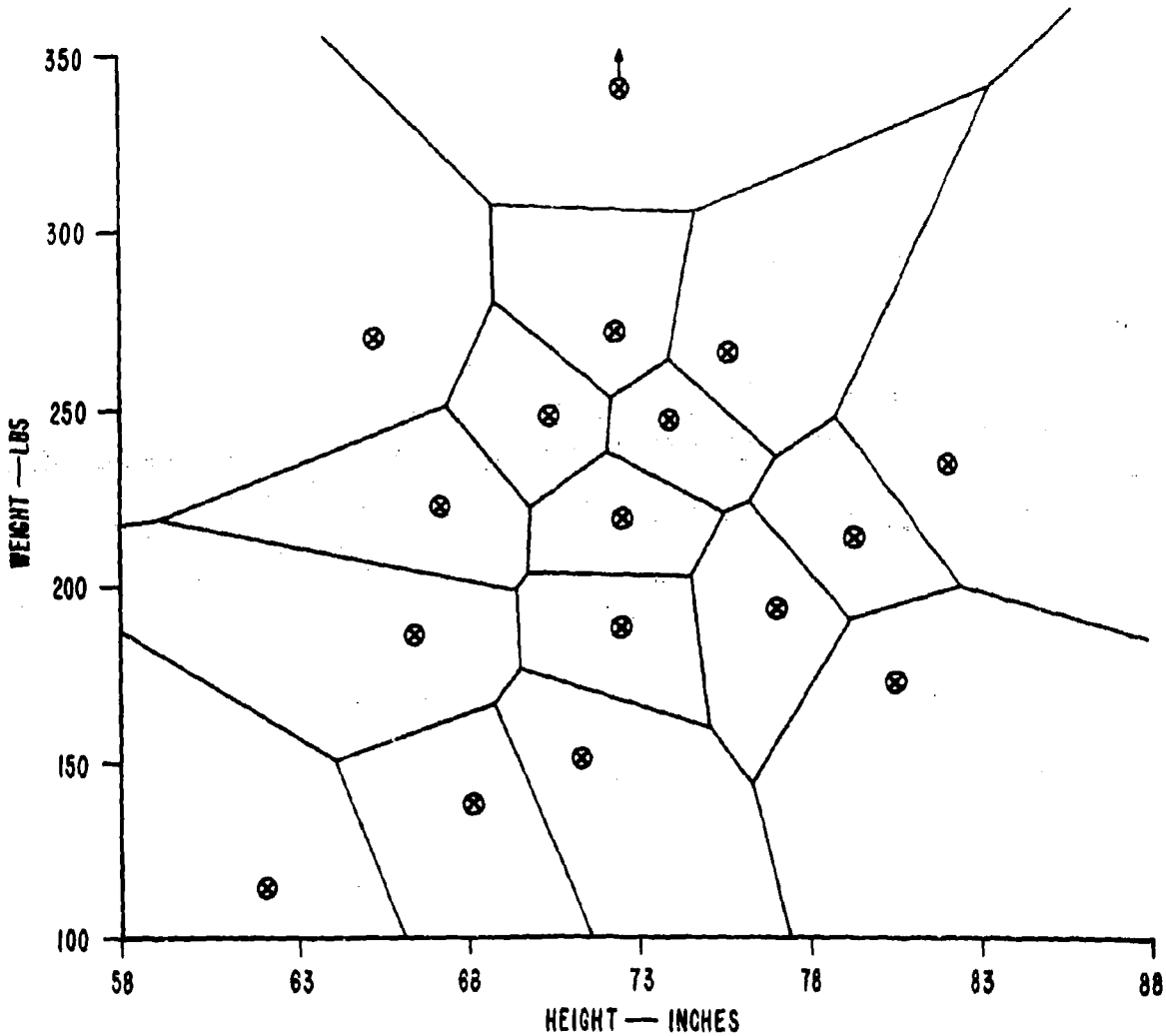


FIG. 16 THE PARTITION FOR ITERATION 5

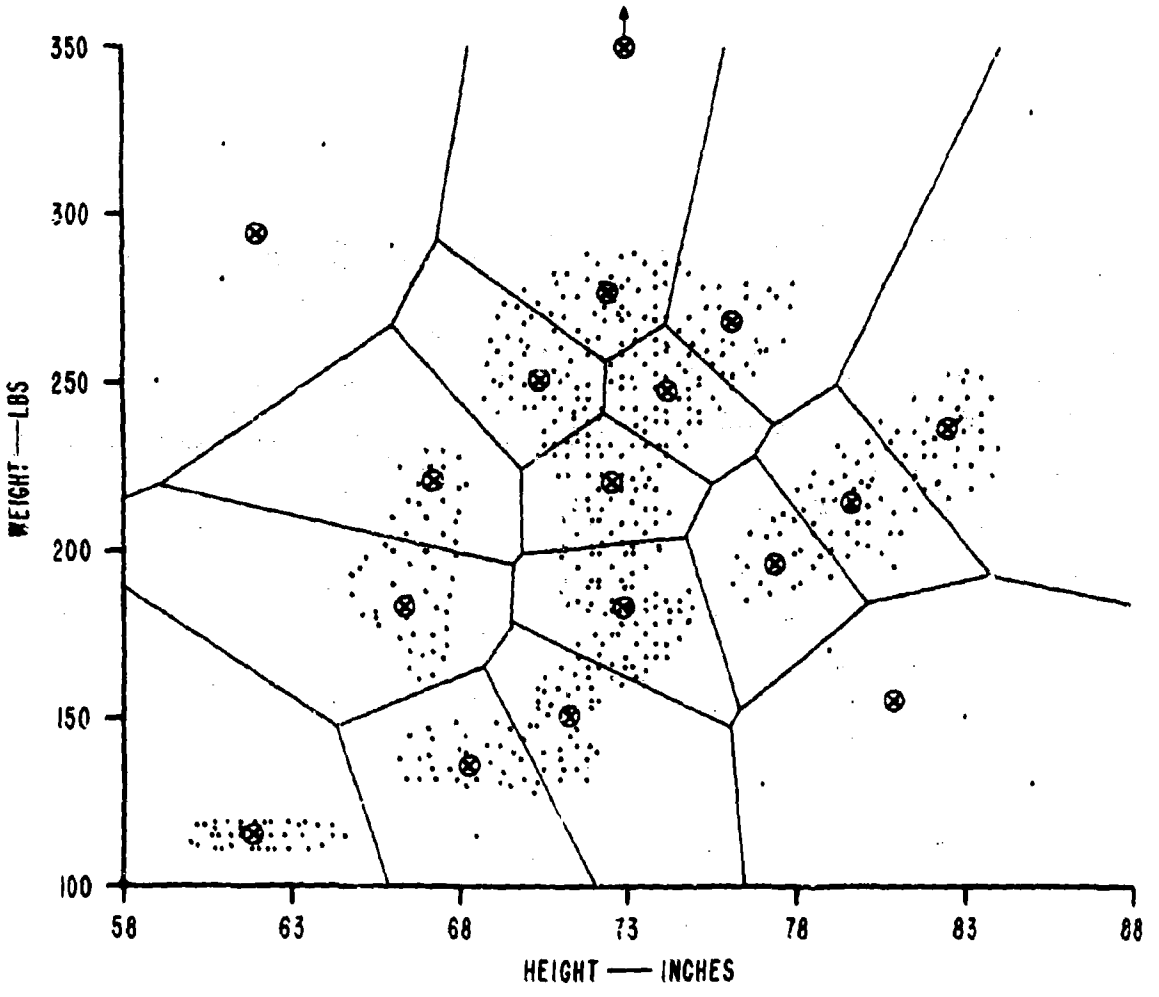


FIG. 17 THE AVERAGE POINTS FOR THE SUBSETS OF ITERATION 5

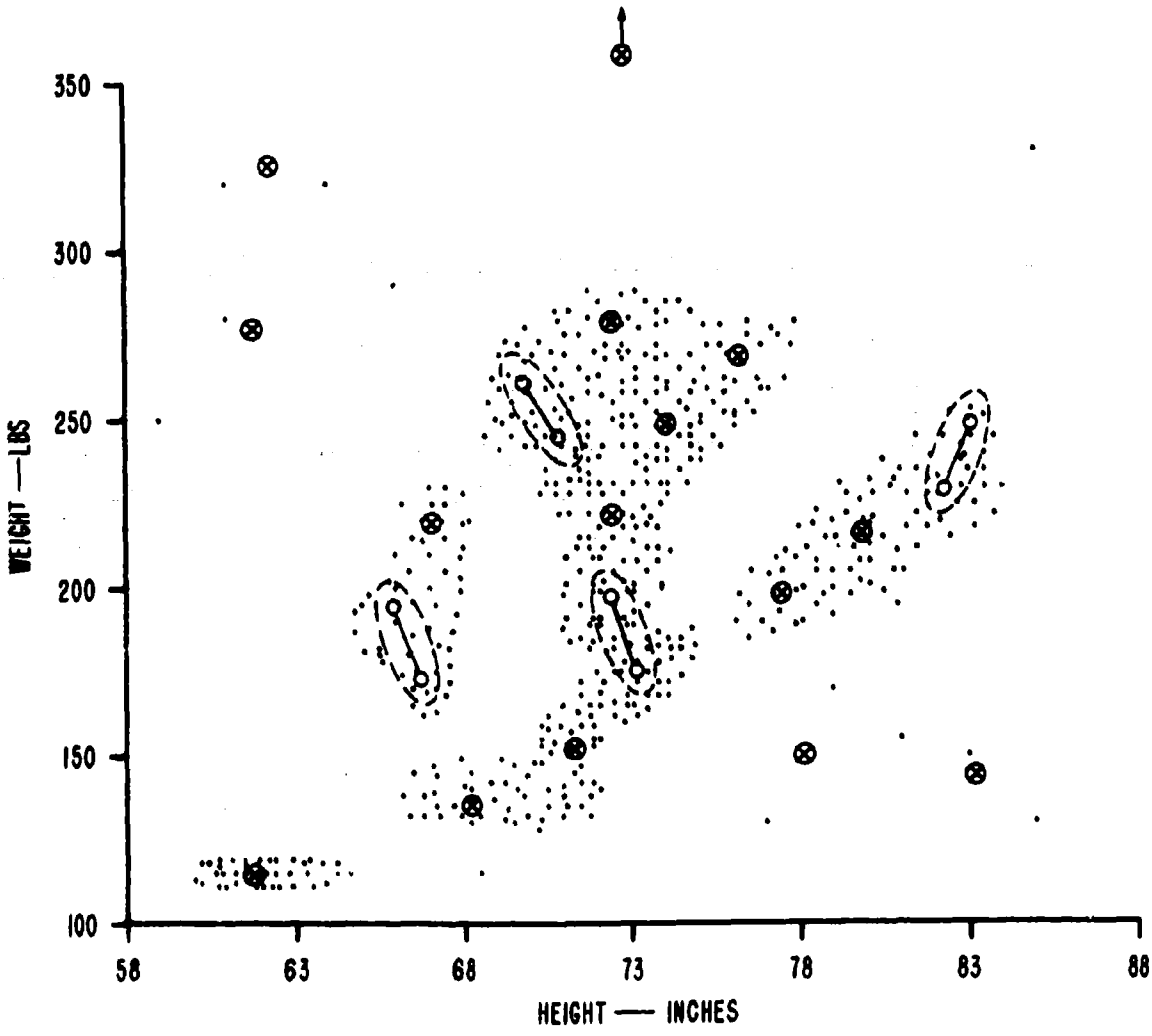


FIG. 18 THE AVERAGE POINTS FOR ITERATION 6

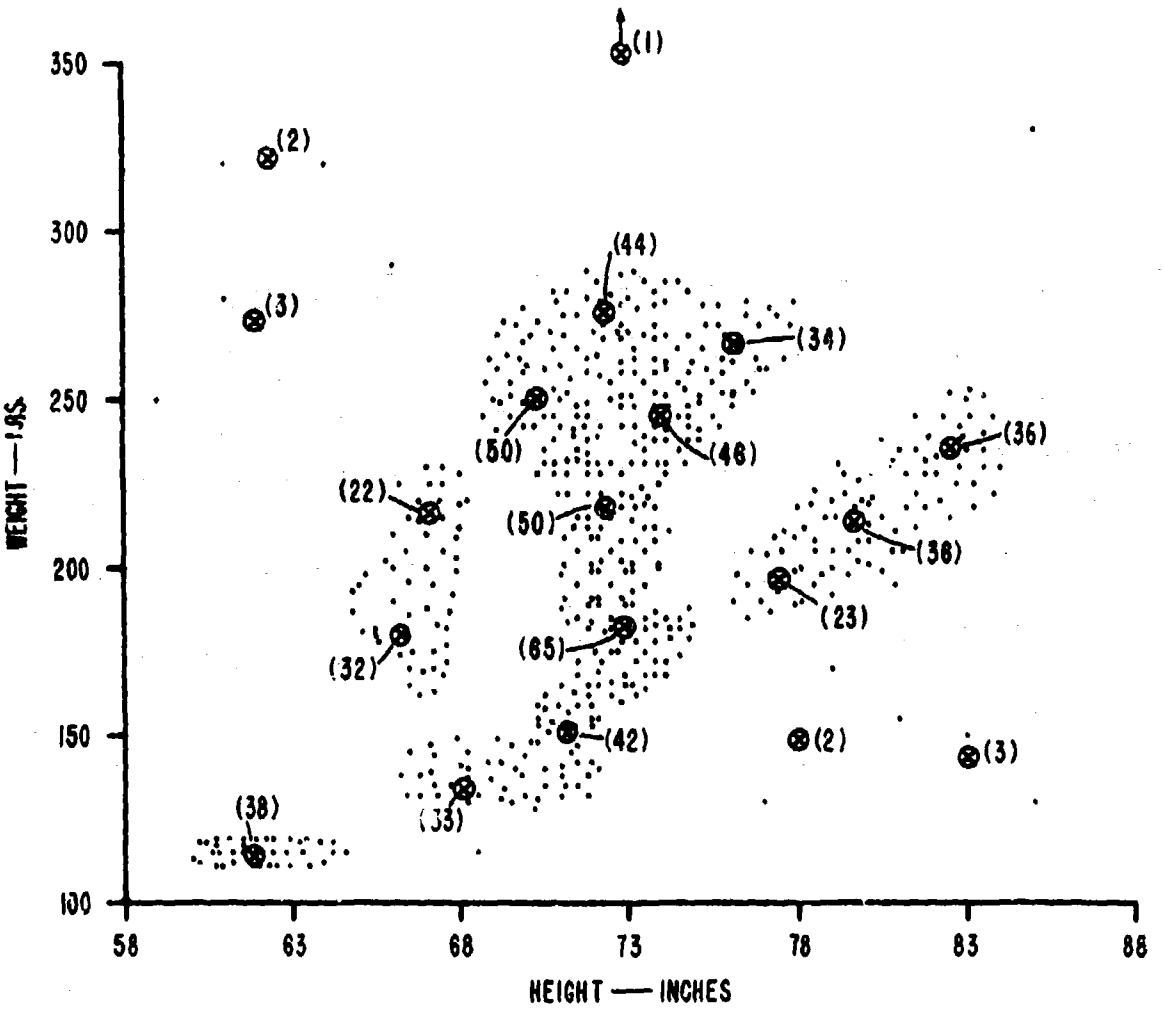


FIG. 19 THE FINAL AVERAGE POINTS AFTER SEVERAL ITERATIONS

D. Mathematical Description

The details of the calculations made in the existing ISODATA-POINTS computer program* are given in this section.

In Fig. 20 we show a computational flow chart of the technique. A glossary of the symbols used in the mathematical description is given next in order to ease the struggle with new notation. Following the glossary we explicitly write the mathematical expression in the sequence calculated for each significant computation made by the program.

Readers not interested in the details of the computation made by ISODATA-POINTS can skip this section (i.e. turn to page 42) with out serious lose.

---

*The processing time for this program is about

$[2.7 \times 10^{-4} \times$ (number of patterns) x (number of cluster points) x (number of dimensions)] seconds/iteration

on the B-5500 computer (at $180/hr) at SRI. The program is written in Algol 60.

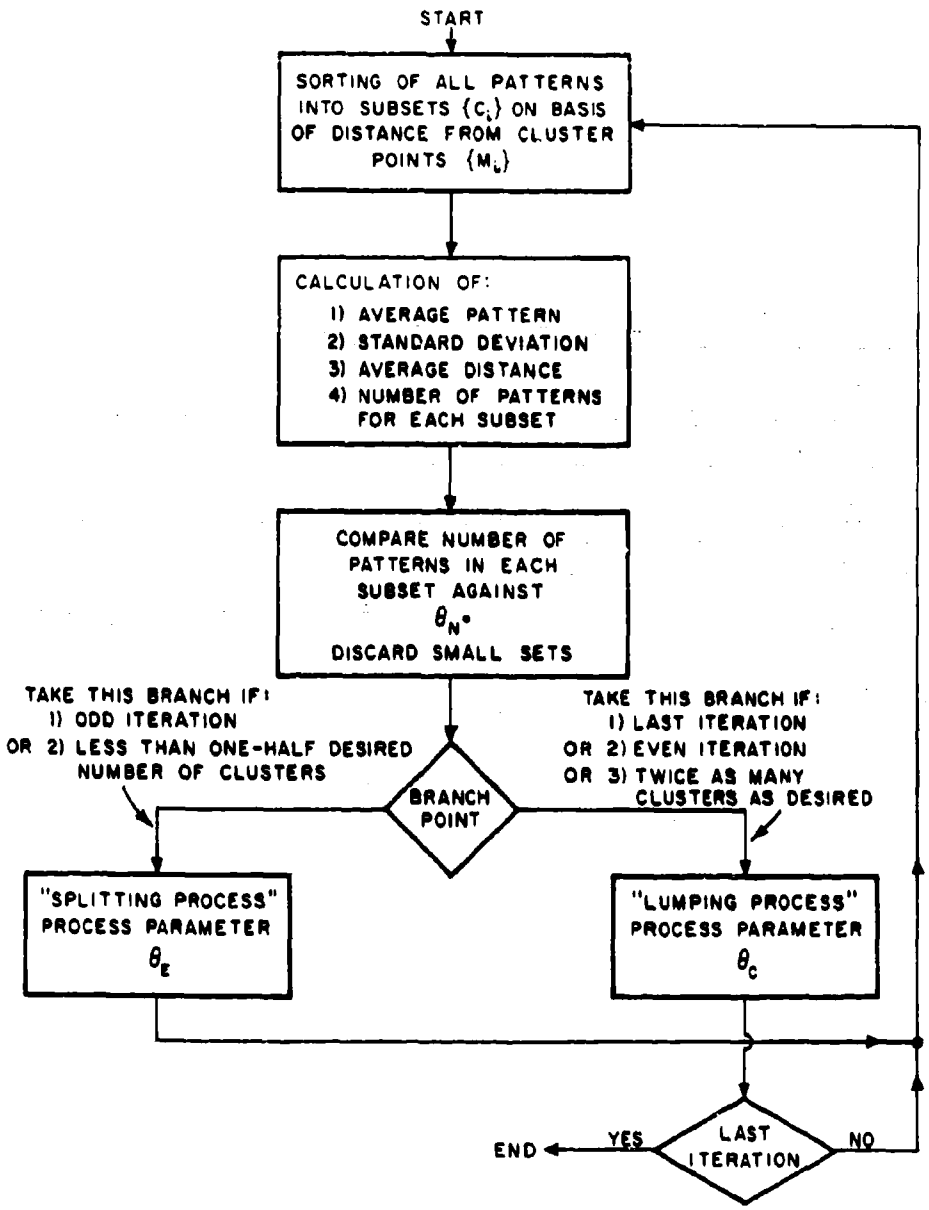


FIG. 20 A FLOW DIAGRAM SHOWING THE COMPUTATIONAL CYCLE OF ISODATA-POINTS

SYMBOLS USED IN MATHEMATICAL DESCRIPTION

SYMBOL	STANDS FOR	FOUND IN STEP NOS.
$\overline{AD}$	Overall average distance of patterns from the average vector of the cluster to which they are assigned. $\overline{AD} = \frac{1}{N} \sum_{i=1}^{NCLUS} (AVEDST_i) \times N_i$	6, 10
$AVEDST_i$	The average distance of the patterns in cluster C_i from the average vector (average point) of that cluster $AVEDST_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \sqrt{(P_j - {}_i\overline{P}) \cdot (P_j - {}_i\overline{P})}$ for all $P_j \epsilon C_i$	4, 6, 10
C_i	The i^{th} cluster	2, 3, 4, 5, 7
D	The dimension (number of components) of a pattern vector.	5, 9
δ_{ij}	The Euclidean distance between the average vector ${}_i\overline{P}$ for cluster C_i and the average vector ${}_j\overline{P}$ for cluster C_j $\delta_{ij} = \sqrt{({}_i\overline{P} - {}_j\overline{P}) \cdot ({}_i\overline{P} - {}_j\overline{P})}$	12
L	The maximum number of pairs of average vectors that can be lumped at one time.	12, 13
M_i	The cluster point (vector) for the i^{th} cluster	1, 2, 11, 14

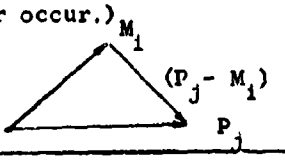
SYMBOL	STANDS FOR	FOUND IN STEP NOS.
N	The total number of patterns.	2, 6
N_i	The number of patterns in the i^{th} cluster C_i.	3, 5, 8, 10, 13
NROWS	The total number of clusters. (Stands for the Number of Rows in the matrix having the cluster point vectors as rows, which is an NROWS x D matrix.)	2, 4, 5, 6, 7, 9, 10, 11, 12 14
P_j	The j^{th} Pattern vector.	2,4
$P_{j\ell}$	The ℓ^{th} component of the j^{th} pattern vector P_j.	5
$_i\bar{P}$	The average pattern vector for the i^{th} cluster C_i	3, 4, 10, 12, 13
$_i\bar{P}_\ell$	The ℓ^{th} component of the average pattern vector $_i\bar{P}$ for the i^{th} cluster.	5
$_i\bar{P}^+$	The positively "split" part of the average vector $_i\bar{P}$. (See Step 10.)	10
$_i\bar{P}^-$	The negatively "split" part of the average vector $_i\bar{P}$. (See Step 10.)	10
$\sigma_{i\ell}$	The standard deviation of the i^{th} cluster C_i in the ℓ^{th} component (dimension) $\sigma_{i\ell} = \sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}(P_{j\ell} - {}_i\bar{P}_\ell)^2}$ all $P_j \epsilon C_i$	5, 9

SYMBOL	STANDS FOR	FOUND IN STEP NOS.
$_i\sigma_k$	The largest standard deviation of all of the components of the patterns in cluster C_i. The largest standard deviation occurs in the k^{th} component. $_i\sigma_k = \max_j \{\sigma_{ij}\}$	9, 10
θ_C	The ISODATA process parameter against which the distance δ_{ij} between pairs of patterns is compared. It controls the "lumping" process. It is supplied by the researcher.	12
θ_E	The ISODATA process parameter against which the maximum standard deviation $_i\sigma_k$ is compared. It controls the splitting process. It is supplied by the researcher.	10
θ_N	The ISODATA process parameter against which the number N_i of patterns in a cluster is compared. It is supplied by the researcher.	7, 10

STATEMENT OF THE GOAL

Given a set of pattern vectors $\{P_j, j = 1, \ldots, N\}$ of dimension D, the goal of ISODATA-POINTS is to sort them into subsets C_i each having N_i members and having small within-group variance, i.e., find a set of average vectors such that each of these average vectors adequately describes that set of patterns lying closest to it. (Measures and criteria for determining when an adequate description of the data has been obtained are being sought.)

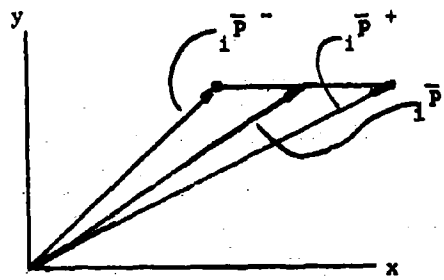
The following steps describe in symbols the manipulations shown graphically in the figures accompanying the two-dimensional example illustrating ISODATA-POINTS.

Step	Computation
1.	Select arbitrary subset of patterns having NROWS elements. These should be chosen as intelligently as possible--i.e., if possible, one from each known sub-class or cluster. Create a set of points that are duplicates of this randomly selected set of patterns. Call this duplicate set the initial "cluster points," = $\{M_i, i=1,\ldots, NROWS\}$
2.	Do for all $j=1,\ldots,N$:[#] For each pattern P_j find M_{i*} such that $(P_j - M_i^*)\cdot(P_j - M_i^*) = \min_i [(P_j - M_i)\cdot(P_j - M_i)]$ where the dot product $A\cdot B = (a_1,\ldots,a_D)\cdot(b_1,\ldots,b_D)$ $= \sum_{i=1}^{D} a_i b_i$ Assign P_j to subset C_i^*. This step sorts the $\{P_j\}$ into subsets on the basis of distance from the $\{M_i\}$ (see figure below). (Ties are arbitrarily decided. They almost never occur.) 

[#]Note that the M_i are not changed during this calculation over all j.

Step	Computation
3.	Compute for all of the i clusters C_i the average vector ${}_i\bar{P}$ of each cluster i $${}_i\bar{P} = \frac{1}{N_i} \sum_{P_j \in C_i} P_j,$$ where N_i is the number of patterns in cluster C_i.
NOTE:	The ΣP_j is obtained as patterns are being sorted.
4.	For all i, i=1,..., NROWS. Compute the average distance $AVEDST_i$ of patterns in cluster C_i from the average vector, of that cluster. $$AVEDST_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \sqrt{(P_j - {}_i\bar{P}) \cdot (P_j - {}_i\bar{P})}$$ For all $P_j \in C_i$
5.	For all i, i=1,...,NROWS, and for all ℓ, ℓ=1,...,D find the standard deviation $\sigma_{i\ell}$ of the i^{th} subset for ℓ^{th} measurement where $$\sigma_{i\ell} = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (P_{j\ell} - {}_i\bar{P}_\ell)^2}$$ For all $P_j \in C_i$
6.	Compute average distance overall, $\overline{AD}$, where $$\overline{AD} = \frac{1}{N} \sum_{i=1}^{NROWS} (AVEDST_i) \times N_i$$ This is the average distance of the patterns from their closest cluster point.

Step	Computation
NOTE:	Certain parameters, θ_N, θ_E, θ_C, L and the total number of iterations, which will be mentioned in the following steps, are provided the program by the researcher. For Steps 7-10 and Steps 12-13 no use is made of the individual patterns. All calculations are made based on means, standard deviations, $AVEDST_i$, and $\overline{AD}$, and the process parameters.
7.	For all i, i=1,..., NROWS. If $N_i < \theta_N$ then discard the i^{th} cluster, C_i, and reduce the number of clusters by 1.
8.	(a) If this iteration is the last iteration, set $\theta_C = 0$ and skip to step 12. (b) If number of clusters is less than or equal to $\frac{1}{2}$ the number desired, then skip the remaining step in 8 and do step 9. (c) If this iteration is an even-numbered iteration, or if number of clusters is greater than or equal to twice the number desired, then skip to step 12.
NOTE:	Steps 9 through 11 comprise the so-called "splitting" process.
9.	For all i, i=1,..., NROWS, for j=1,...,D., find k and $i\sigma_k$ such that $$i\sigma_k = \max_j(\sigma_{ij})$$
10.	For each i, i=1,...,NROWS (a) If $i\sigma_k > \theta_E$ and ((if $AVEDST_i > \overline{AD}$ and if $N_i > 2\theta_N + 2$) or (NROWS ≤ 0.5 x number of clusters desired)) then create

Step	Computation
10. (Cont.)	${}_i\bar{P}^+ = {}_i\bar{P} + (0,\ldots,0,+1,0,\ldots,0)$ (k^{th} component) The +1 is placed in this k^{th} component element (corresponding to ${}_i\sigma_k$ for cluster i) in order to split the cluster along the direction having the maximum variation. ${}_i\bar{P}^- = {}_i\bar{P} + (0,\ldots,0,-1,\ldots,0)$ (k^{th} element) (b) ${}_i\bar{P}$ is replaced by ${}_i\bar{P}^+$, and ${}_i\bar{P}^-$ is added to the list of average vectors, which increases the number of clusters by 1. (See figure below) 
11.	Start the process again at Step 2. Use the $\{{}_i\bar{P}, i=1,\ldots,NROWS'\}$ as the new set $\{M_i'\}$ in place of the existing set $\{M_i\}$. (NROWS' is the number of clusters <u>after</u> splitting or lumping.)
NOTE:	Steps 12 through 14 comprise the so-called "lumping" process.
12.	For all i, i=1,..., NROWS: For all j > i, j=i + 1,..., NROWS: (1) Compute the pairwise distance δ_{ij} between average points where $\delta_{ij} = \sqrt{({}_i\bar{P} - {}_j\bar{P})\cdot({}_i\bar{P} - {}_j\bar{P})}$

Step	Computation
12. (Cont.) (2)	If $\delta_{ij} < \theta_C$ then place $\delta_{i_\ell j_\ell}$, i_ℓ, j_ℓ in an ordered (3 x L) matrix.

$$\begin{bmatrix} \delta_{ij} & \delta_{i_2 j_2} & \cdots & \delta_{i_L j_L} \\ i_1 & i_2 & \cdots & i_L \\ j_1 & j_2 & \cdots & j_L \end{bmatrix}$$

where $\delta_{i_1 j_1} < \delta_{i_2 j_2} \ldots < \delta_{i_L j_L}$

NOTE: L (which is ≤ 9 for programming convenience) controls the maximum number of pairs of clusters that are lumped together.

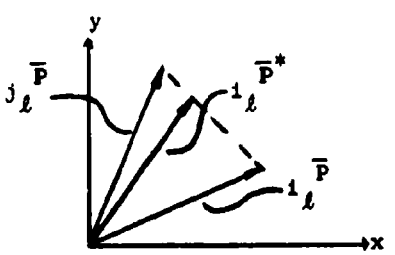
13. For all ℓ, $\ell = 1, \ldots, L$.

If ${}_{i_\ell}\bar{P}$ and ${}_{j_\ell}\bar{P}$ have not been previously used in lumping, then

(1) Compute

$${}_{i_\ell}\bar{P}^* = \frac{1}{N_{i_\ell} + N_{j_\ell}} \left[N_{i_\ell} \left({}_{i_\ell}\bar{P} \right) + N_{j_\ell} \left({}_{j_\ell}\bar{P} \right) \right]$$

(2) Replace ${}_{i_\ell}\bar{P}$ with ${}_{i_\ell}\bar{P}^*$ and delete ${}_{j_\ell}\bar{P}$ from the list of average vectors (reducing the number of clusters by 1).



14. If more iterations are to be done (this is at discretion of investigator), start the process again at Step 2. Use the $\{{}_i\bar{P}, i=1,\ldots \text{NROWS}'\}$ as the new set $\{M_i'\}$ in place of the existing set $\{M_i\}$. If this was the last iteration, then end process.

E. Analysis of the Height vs. Weight Data Using Principal Components Analysis

An alternative method of describing and analyzing the data of Section C (the two-dimensional example) would be principal components analysis.

It may be objected that principal components analysis should not be applied to data that is as heterogeneous as the data in this two-dimensional example. We agree. At least part of our point is that it is not easy in high dimensions to determine just how heterogeneous the data is.

"The (Principal Component Method) is a relatively straightforward way of 'breaking down' a covariance or correlation matrix into a set of orthogonal components or axes equal in number to the number of variates concerned. These correspond to the latent roots and the accompanying latent vectors....of the matrix. The method has the property that the roots are extracted in descending order of magnitude, which is important if only a few of the components are to be used in summarizing the data. The vectors are mutually orthogonal, and the components derived from them are uncorrelated."[3] The greatest possible "scatter" of n points projected onto a given number s of coordinate axes in a k-dimensional space ($s \leq k$) is obtained by this method.[4]

The average point of the height vs. weight data is (724,208), (the inches are multiplied by 10) and the covariance matrix is

$$\begin{bmatrix} 2667 & 1111 \\ 1111 & 2814 \end{bmatrix}$$

The first eigenvalue is 3854 and the corresponding eigenvector is (.936,1.00). The second eigenvalue is 1627 and the corresponding eigenvector is (1.00, -.936). In Fig. 21 we have plotted these eigenvectors as a second set of "coordinate axes" with the mean value of all of the data as origin. The length of the vectors is proportional to the magnitude of the associated eigenvalue.

The direction of the first eigenvector indicates that generally there is a positive correlation between height and weight; that is, weight increases with height. This "accounts for" about 70% of the variance. Both height and weight contributed about equally to this component.

The second eigenvector displays the extent to which height and weight are negatively correlated. Again both height and weight contribute about equally to this component.

---

*The exact values of "scatter" for n data points in k dimensions are given by Wilks.[5]

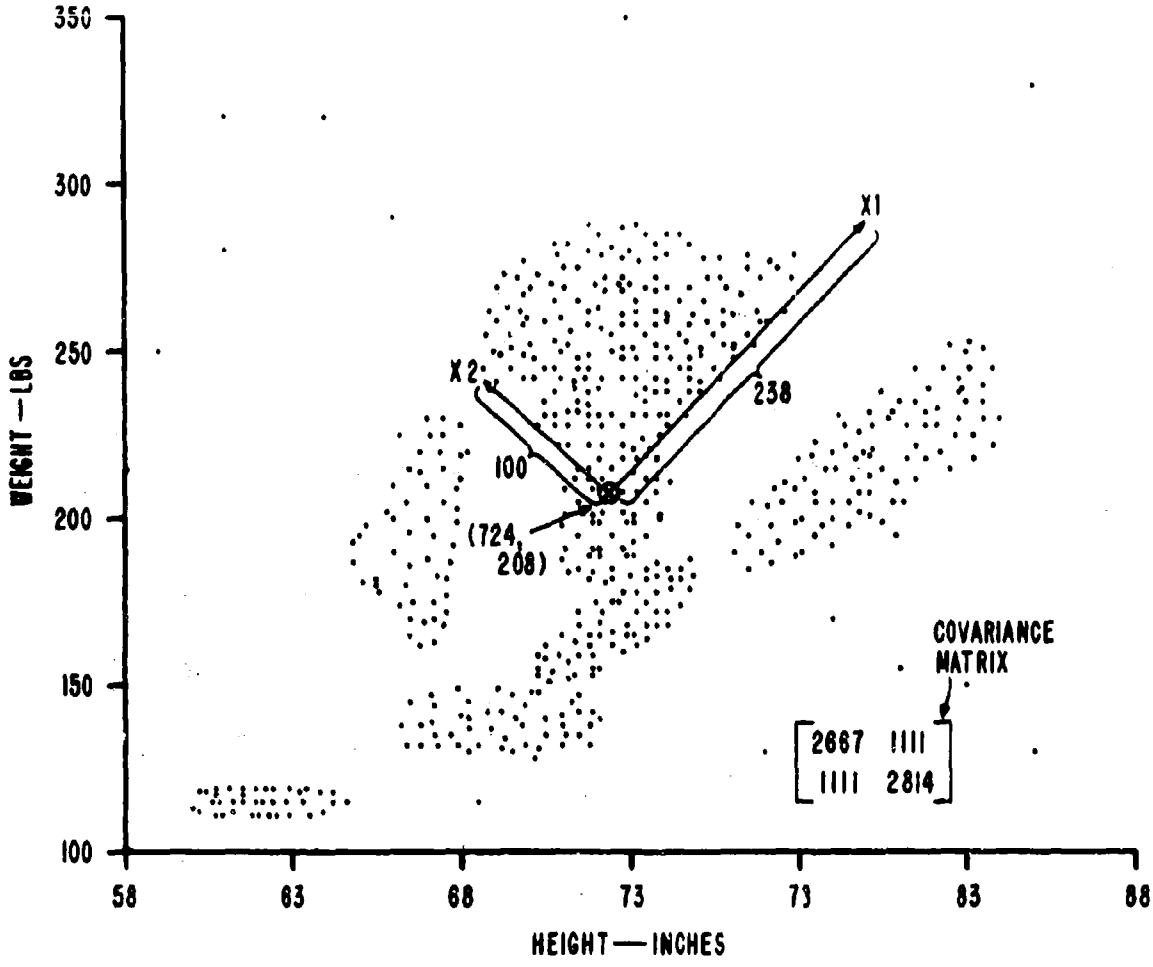


FIG. 21 THE PRINCIPAL COMPONENTS OF THE WEIGHT vs. HEIGHT DATA OF THE TWO-DIMENSIONAL ILLUSTRATIVE EXAMPLE

These descriptions relate primarily to directions in the data. The largest eigenvector gives the direction along which a scale should be set up to get maximum variation in the data. If the goal of the data analysis is to find such a scale, then this is a very reasonable analysis technique.

However, such "direction-finding" techniques tend to ignore details in the data such as the existence of isolated clusters, e.g., the cluster around the point (62", 115 lbs.) in Fig. 21.

Clustering techniques like ISODATA-POINTS ignore direction while clustering the data in the pattern space. However, the average points obtained can be used to derive the directional characteristics of the data. They are primarily sensitive to the density of the patterns in pattern space. They are well suited to "zooming" in on the detailed structure of the data. They also can serve as methods for a preliminary sorting of patterns into relatively homogeneous subsets for further statistical processing. (This sorting can prevent the confounding of two disparate effects resulting from treating these effects as if they were the result of the same (simple) cause.)

It may also be objected that ISODATA-POINTS has a certain arbitrariness about it and that by setting the process parameters differently we would obtain different average points. It is true that different average points can be obtained by varying the process parameters. However the results of the clustering plus specification of the cluster parameters used, provides an objective and useful description of the data.

In complex data we have found that there are a variety of valid clusterings depending on the number of average points used, on scaling, and on the structure of the data itself. For example, if the data consists of tight clusters of data whose distances apart are large with respect to the "diameter" of a cluster then the number of clusters will not vary even with wide variations in the process parameters. If, on the other hand, the data is uniformly distributed in pattern space, then the number of clusters found will tend to vary rather smoothly with changes in the process parameters. The way that the number of clusters varies as a function of the process parameters can be used to describe the structure of the data.

For these reasons, we feel that what must seem arbitrariness to some is a flexibility that is needed for the analysis of real data. We feel that this flexibility is not detrimental in the case of data analysis by clustering.

We agree with John Tukey's appeal for good judgment in place of rigorous optimization:

> "Scientists know that they will sometimes be wrong; they try not to err too often, but they accept some insecurity as the price of wider scope. Data analysts must do the same."[16]

And further.

> "If data analysis is to be well done, much of it must be a matter of judgment, and 'theory,' whether statistical or non-statistical, will have to guide, not command."[17]

And finally, we quote from Tukey at some length, because of the relevance of his remarks to clustering techniques.

> "Practicing data analysis". If data analysis is to be helpful and useful, it must be practiced. There are many ways in which it can be used, some good and some evil. Laying aside unethical practices, one of the most dangerous (as I have argued elsewhere (Tukey, 1961b)) is the use of formal data-analytical procedures for sanctification, for the preservation of conclusions from all criticism, for the granting of an imprimatur. While statisticians have contributed to this misuse, their share has been small. There is a corresponding danger for data analysis, particularly in its statistical aspects. This is the view that all statisticians should treat a given set of data in the same way, just as all British admirals, in the days of sail, maneuvered in accord with the same principles. The admirals could not communicate with one another, and a single basic doctrine was essential to coordinated and effective action. Today statisticians can communicate with one another, and have more to gain by using special knowledge (subject matter or methodological) and flexibility of attack than they have to lose by not all behaving alike.
>
> In general, the best account of current statistical thinking and practice is to be found in the printed discussions in the Journal of the Royal Statistical Society. While reviewing some of these lately, I was surprised, and a little shocked to find the following:
>
> 'I should like to give a word of warning concerning the approach to tests of significance adopted in this paper. It is very easy to devise different tests which, on the average, have similar properties, i.e., they behave satisfactorily when the null hypothesis is true and have approximately the same power of detecting departures from that hypothesis. Two such tests may, however, give very different results when applied to a given set of data. This situation leads to a good deal of contention amongst statisticians and much discredit of the science of statistics. The appalling position can easily arise in which one can get any answer one wants if only one goes around to a large enough number of statisticians.'[18]

> To my mind this quotation, if taken very much more seriously than I presume it to have been meant, nearly typifies a picture of statistics as a monolithic, authoritarian structure designed to produce the 'official' results. While the possibility of development in this direction is a real danger to data analysis, I find it hard to believe that this danger is as great as that posed by over-emphasis on optimization.
>
> Facing uncertainty: The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: 'Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.' Data analysis must progress by approximate answers, at best, since its knowledge of what the problem really is will at best be approximate. It would be a mistake not to face up to this fact, for by denying it, we would deny ourselves the use of a great body of approximate knowledge, as well as failing to maintain alertness to the possible importance in each particular instance of particular ways in which our knowledge is incomplete."[19]

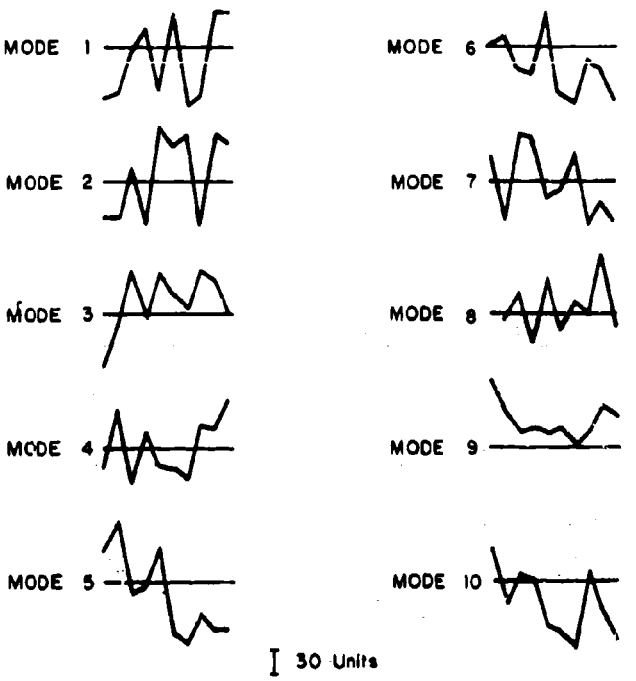
We are presently investigating in more detail the relationships between clustering and "direction-finding" techniques. It appears at this time as though they are qualitatively different, and that they should be used to complement each other in data analysis.

V. EXPERIMENTAL RESULTS FROM COMPUTER IMPLEMENTATION OF ISODATA-POINTS

In order to understand and evaluate the technique, we have performed a series of experiments. These experiments have been of three types: those designed to validate ISODATA-POINTS, those designed to illustrate graphically the approach taken, and those designed to analyze data from the real world.

The detailed results that we have obtained using the ISODATA-POINTS program will be contained in a larger report. We shall not attempt to repeat the contents of that report, but rather shall extract some of the results that we consider particularly significant.

The validation experiments were constructed from data whose structure was well known, in order that we might evaluate the clustering obtained by ISODATA-POINTS. The data was obtained by adding Gaussian random noise to 10 prototype patterns which serve as the ideal versions of the noisy vectors. Each vector had 10 analog dimensions and can be displayed as shown in Fig. 22(a). The values in each dimension were coded into a 10-bit binary number using "snake-in-the-box" codes



(a)

(b)

FIG. 22(a) TEN WAVEFORMS REPRESENTING THE TEN-DIMENSIONAL PROTOTYPE PATTERNS
(b) TWO NEARLY IDENTICAL PROTOTYPE WAVEFORMS

(a modified gray scale coding). These 10 dimensions were then combined to give a 100-bit pattern. The closest intermean distance was 108 units of the original 10-dimensional space, or roughly 20 bits apart in Hamming distance. The covariance matrix was the same for all distributions and was the scalar matrix, σI, I being the identity matrix, and σ=30 units, where σ is the standard deviation of the distributions. The size of σ=30 is also indicated in Fig. 22(a).

The patterns were processed by the ISODATA-POINTS program without specifying the distributions from which they came. The process parameters were varied by the authors until the program sorted the patterns into 11 clusters. We assigned each cluster to the mode whose patterns predominated in that cluster. Doing this the program classified 98% of the patterns correctly. The Bayesian decision-theoretic optimum separating planes, which were positioned using a priori knowledge of the location of the means, achieved a 99% correct rate using unquantized data.

In a later experiment only two distributions were used. The waveforms of the two means are shown in Fig. 22. Their means were but 56 units (about 10 bits) apart, while they still had a standard deviation of about 30 units (about 6 bits). We obtained estimates of the mean values, again without knowledge of the pattern mode from which the data came. The values obtained were only slightly different from the correct means. The optimum decision plane gave a percentage correct classification of 81% while ISODATA-POINTS (using two large clusters and two quite small ones) obtained 78%.

In a second validation experiment we constructed 48 pattern vectors. half of which had the last six bits + 1's and half of which had the last six bits - 1's. The first 24 bits of these 30-bit vectors were then filled with pattern vectors positioned so as to have all pairwise Hamming distances between patterns exactly 24 bits in this 24-dimensional space. The ISODATA-POINTS program proved capable of extracting the six consistent bits of this 30-bit vector--disregarding the rest of the "noise" for which bits were not consistent. The program was 100% correct in its classification of these vectors. Again no categorization information was used to cluster the patterns.

In another experiment designed primarily to illustrate ISODATA-POINTS graphically, we drew a set of O's and a set of Q's on a 10 x 10 retina of squares. The O's and Q's had no registration noise (i.e., were not translated or rotated) but did have bits of noise added randomly, adjacent to the main outline of the O's and Q's. The program proved capable of dividing the O's from the Q's perfectly. In addition, the subtracting of the average "O" pattern from the average "Q" pattern emphasized the fact that the tail of the Q's was the primary distinction between these two classes of patterns.

We felt that the utility of ISODATA-POINTS would be most clearly indicated by application to data drawn from actual research situations. We chose two such situations. The first was the sorting of a set of sociological questionnaires, relating to the attitude of 209 scientists at various Air Force laboratories. In this instance it was somewhat difficult to specify a reasonable sorting of the questionnaires into groups or categories. The second was a set of weather data, relating to ceiling height prediction in Washington, D.C. In this situation, we could perform a preliminary sort (i.e., define a classification) using the ceiling height that actually did occur.

The sociological data were obtained from the Systems Analysis Laboratory of the Management Sciences Division of Stanford Research Institute.* We found the groups into which the ISODATA-POINTS technique divided the scientists' questionnaires had reasonable internal consistency as measured by the mean deviation from the average point. Conversations with the research sociologists have indicated that these groupings are meaningful in terms of their experience with the personnel in the laboratories. We were even able to obtain distances (in terms of the measurements made by the questionnaire) between the average points of these groups and to obtain the spatial relationships of the groups in three dimensions. We feel that such groupings can point out the characteristics of the people to whom a questionnaire is given. These characteristics appear to be useful in revising the questionnaire for future use.

The ceiling height weather data provided us with the opportunity to investigate three aspects of the ISODATA-POINTS process. In this sense the weather problem is a very "rich" problem suggestive of many useful modifications of ISODATA-POINTS. In particular it allowed us:

(1) To evaluate its ability to predict ceiling height;

(2) To evaluate its capabilities for measurement selection**;

(3) To learn how ISODATA-POINTS exhibits the structure of experimentally-obtained data.

The performance obtained by the technique was slightly better than persistence forecasting. (Persistence forecasting is the technique of forecasting that predicts that the same conditions that exist at the present time will be in effect at some later time--(in this case five hours). This prediction was, however, made without utilizing the categorization information. In the near future we hope to improve the prediction score by utilizing the actual ceiling height that occurred five hours later for a preliminary sorting of the patterns.

---

* The analysis of this data was supported by contracts from the Behavioral Science Division of the Air Force Office of Scientific Research.

** By measurement selection we mean the determination of those predictors or measurements that are most useful in discriminating between classes.

Utilization of the typical patterns, or cluster centers, as we have called them, allows us to evaluate the measurements that define each of the patterns. We have done this for the weather data and have obtained agreement with both the intuitive ratings of a meteorologist and measurement evaluations obtained by statistical techniques. We have been able to go somewhat further than this in one respect. From our examination of the data it is evident that when one considers only the weather records that resulted in low ceiling heights, the important measurements are different from those considered important when one uses high ceiling height records as well. This indicated to us that predictors or measurements that may be essential in one region of pattern space need not be even useful at all in other regions. Any over-all statement regarding predictor worth that averages together performance in different regions of pattern space seems destined to obscure such important details.

By plotting the cluster centers in a plane using the distances of the clusters from each other, we were able to see the structure of this experimental data. This we found quite suggestive of new measurements that should be made.

VI HOW THE OUTPUT FROM AN ISODATA-POINTS ANALYSIS CAN BE USED

The information supplied by an ISODATA-POINTS clustering consists of:

1. For each cluster:

   a) The number of patterns in it;

   b) The average distance of the patterns in that cluster from the average point of the cluster;

   c) The number of patterns from each class that are in this cluster;

   d) The identity of the patterns that are in that cluster.

2. The positions of a set of average points that the process has located in regions of high pattern density, and the standard deviation of the patterns around these average points for each of the pattern components.

3. The distance of all patterns from all of the final average points.

4. The distance of each average point from every other average point, i.e., the distances between all possible pairs of average points.

5. The average distance (taken over all patterns) from a pattern to its closest cluster point.

Using this information it is possible to learn a great deal about the structure of the patterns in pattern space. The gross structure of the data is obtained by examining the spatial relationships between the average points. Note that the number of average points is small enough to allow comparison of each average point with every other average point. We have found that a most useful way of comparing these average points is a graphical plot. It is not possible to draw the plot in the original pattern space because it has too many dimensions. However, by using the distances between pairs of average points, it is possible to plot at least three average points on a flat surface. We have found that with real data we have frequently been able to plot on a plane more than three average points with sufficient accuracy to aid our intuition. The distribution of patterns around these average points can be plotted using distances from these average points and a more detailed understanding of the fine structure of the data obtained.

Using the information now available in the program some evaluation of the significance of a given clustering is possible. One criterion of clustering that can be used is average distance ($\overline{AD}$) of a pattern from its closest average point (the average point for the cluster to which a pattern belongs). In Fig. 23 we show the value of $\overline{AD}$ vs. the number of clusters for the two-dimensional example as the iterations progressed. Note that after the fourth iteration it changes little as the number of masks changes. This criterion is, however, probably not as sensitive as might be desired. By using the distances between average points as well it appears possible to determine if the patterns are compactly clustered.

We are continuing to seek new ways in which the results of the clustering can be analyzed.

VII SUGGESTIONS FOR FURTHER RESEARCH

The following further research is suggested by the work thus far. In addition to research on algorithms for ISODATA-LINES and ISODATA-PLANES, we intend to investigate:

(1) Criteria for clustering, in order to improve our ability to interpret the results of clustering and to facilitate a more efficient manipulation of the process parameters.

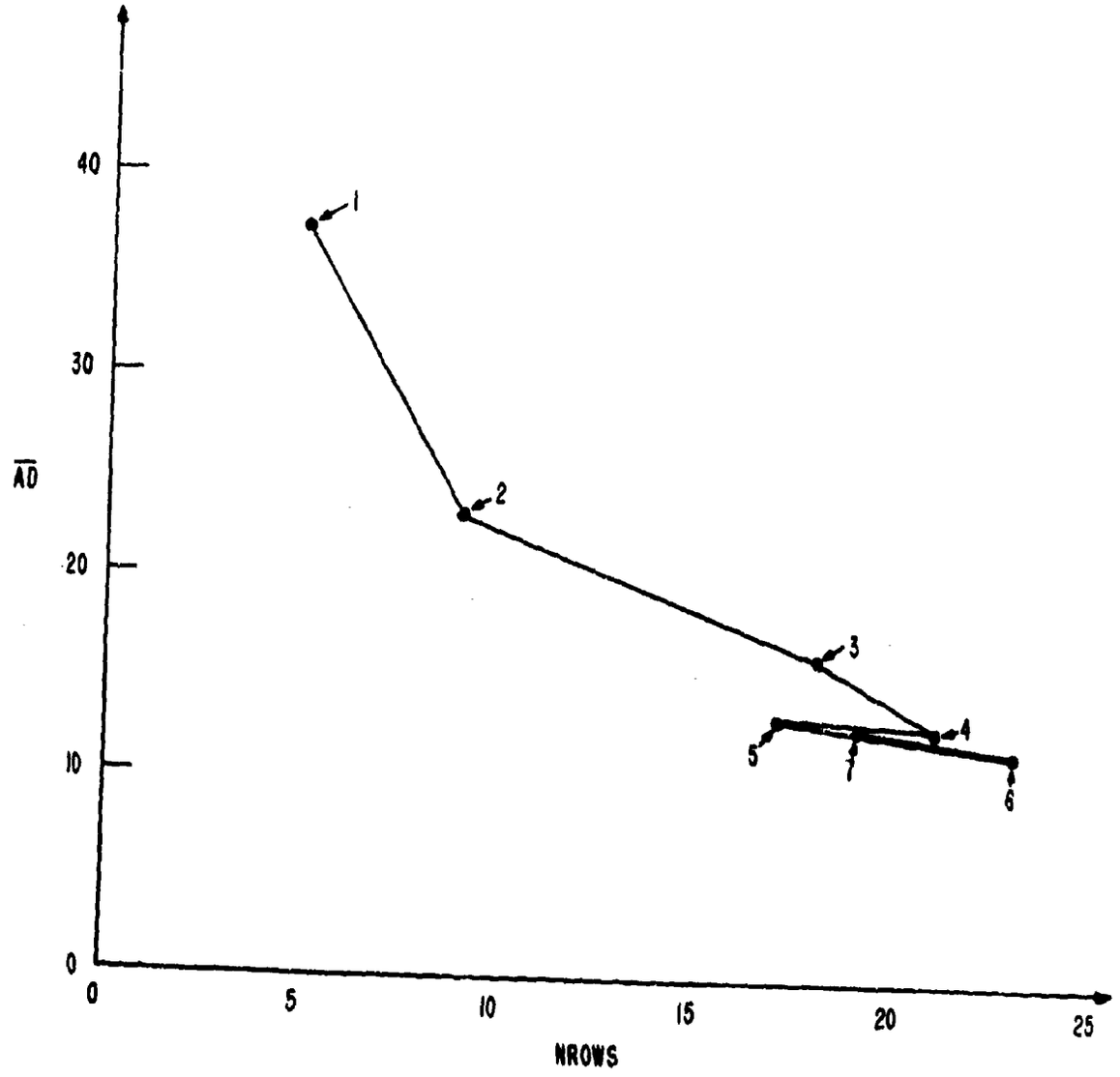


FIG. 23 THE AVERAGE DISTANCE OF A PATTERN FROM ITS CLOSEST CLUSTER POINT vs. THE NUMBER OF CLUSTERS FOR ITERATIONS 1-7

(2) Methodology for using clusters of patterns. Here we seek methods of displaying and analyzing the results of clustering.

(3) Classification techniques based on ISODATA-POINTS, ISODATA-LINES and ISODATA-PLANES that use distance from points, line segments, and planar segments as the criteria for determining the class membership of an unknown pattern.

(4) Actual hardware implementations of any methods that prove promising after thorough investigation by computer programs. These implementations would be used for data analysis and classification on the basis of distance from points, lines, and planes.

(5) Applications using computer programs implementing ISODATA-POINTS, ISODATA-LINES, and ISODATA-PLANES on real-world problems.

In the following five sections we shall discuss these areas for further research in some detail.

1. Criteria for Clustering

So far, in developing the ISODATA techniques we have contented ourselves with using intuitively satisfying criteria in the "decision-making" in the computer program. At this time, we feel that we should investigate additional analytical justification and possibly entirely new criteria. The needed criteria are:

(1) Criteria that could help determine the "goodness of fit" of a given clustering. These criteria would help define "convergence" for ISODATA-like procedures that learn without a teacher.

(2) Criteria for lumping and splitting of the clusters

One important aspect of this part of the work is the determination of the effect of changing the scaling function used for various measurements, e.g., changing from linear to logarithmic scales. This will have an effect on the clusters found. We need to know more about the extent of the effect.

There exist interesting statistical problems in this work. For example, Dr. Charles Dawson of SRI has been able to show that the sum of an infinite set of n-dimensional multivatiate Gaussian distributions having means distributed uniformly along a straight line segment can be considered as n-1 dimensional distributions lying in hyperplanes having the straight line as their normal vector, except very close to the end of the line segment. This model seems an interesting one for the case of detecting a known signal with time-varying amplitude. It also is quite close to a model for one ISODATA-LINES cluster.

2. Methodology for Utilizing Clusters of Patterns

In our work thus far we have developed several methods and programs to aid us in seeing the fine structure of data after clustering. Two important ones are:

(1) The cluster center plot. By using the distance between cluster centers we are able to plot the relative positions of the cluster centers on a plane. We can always plot three such centers and still satisfy the inter-point distance constraints exactly. Frequently we have found it possible to plot more than three on a plane. The exact number that it is possible to plot depends on the spatial relationships that exist in the data.

(2) A distance-from-cluster-center computer program. Using this program we are able to obtain a histogram of the distances of the patterns from the various cluster centers. This gives indications of the distribution of the patterns about the cluster centers--i.e. are they loose or tight clusters, etc. This particular program is useful in setting thresholds and weighting distances between pairs of clusters for the classification of patterns.

It is essential that we develop other methods of rapidly manipulating these clusters of data in order to learn various things about the fine structure of the patterns. We have found that the ideas come most easily in attempting to analyze real data. We are particularly interested in drawing together these techniques to develop a coherent methodology.

3. Classification Using Distance from Lines and Planes

ISODATA-POINTS can be used as a mode-seeking classification technique. Our work on the development of ISODATA-POINTS has also helped us understand so-called piecewise-linear error-correction classification techniques. It therefore seems plausible that the development of ISODATA-LINES and ISODATA-PLANES should help the development of classification techniques that are based on the distance from a set of line segments or from a set of planar segments, where different sets of line segments for example, would be associated with the different classes (see Fig. 24).

For certain classes of patterns, such as speech, word recognition, or speaker recognition, and other non-stationary time-series analysis problems, this type of technique may prove quite powerful.* Another such application might be optical pattern recognition--specifically with

* This seems all the more true when the possibility of optical implementation exists and makes pattern dimensions of 1000 reasonable.

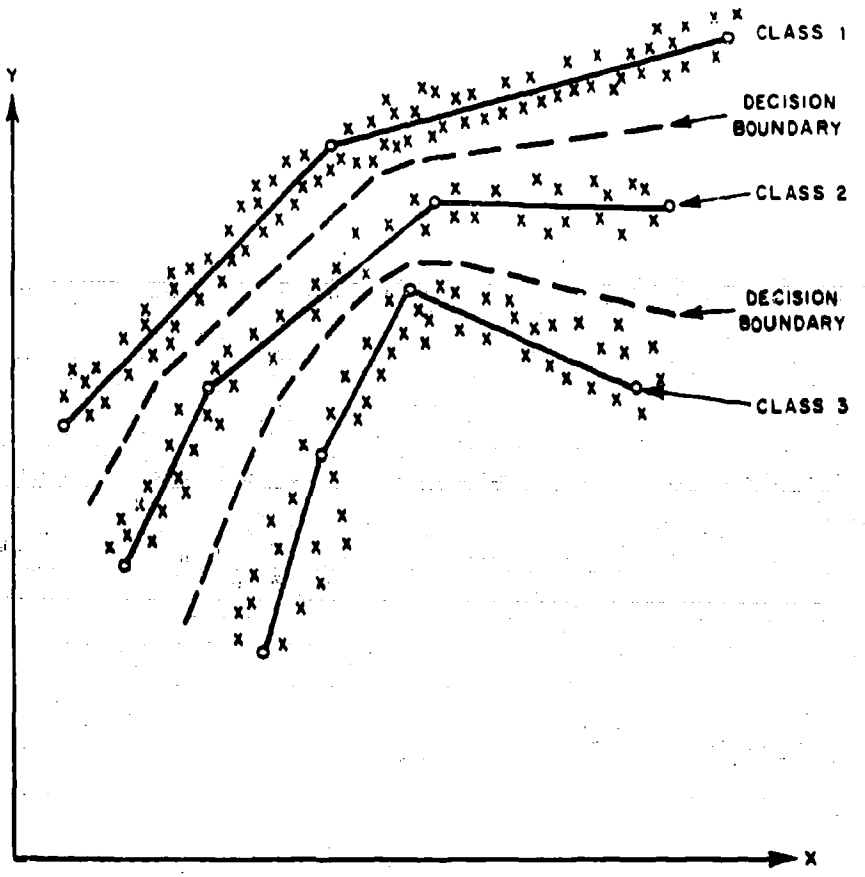


FIG. 24 CLASSIFICATION OF PATTERNS BASED ON DISTANCE FROM PIECEWISE-LINEAR CURVES

respect to recognizing patterns in spite of translation and rotation. We have some evidence that translation of a pattern in one direction creates a straight line in measurement space so long as the rate of change of overlap between the pattern and its translate is a constant. These techniques also lend themselves to the use of a priori probabilities and cost functions.

4. Implementation of ISODATA

In Fig. 25 we show three implementations capable of computing the minimum Euclidean distance from:

(1) An n-dimensional point (the point is specified by a vector M_o).

(2) An n-dimensional line segment.

(3) An n-dimensional planar segment.

These assemblies could therefore be used as basic units for classification using distance from lines or planes. These assemblages are particularly useful when it is not necessary during training to vary the correction factors $M_i \cdots M_j$ (shown in Fig. 25) after each pattern is classified.

The amplifiers shown are variable-gain linear amplifiers.

Both the patterns and the weights can be optical masks in these implementations. If the pattern could be put in the form shown in Fig. 26, this would allow the detailed examination of non-stationary time series with the only requirement being that the system response time not be slower than the sampling rate.

5. Applications

Data usually analyzed by time series analysis, particularly non-stationary series such as speech and business trends, seem to provide suitable sources of data for ISODATA-LINES.

A good source of data for ISODATA-PLANES would be prediction problems where many predictors are used to predict one quantity--for example, ceiling height or air turbulence in Meteorology.

Data from the social sciences would provide excellent data usually suitable for analysis by at least one of the three ISODATA techniques.

The techniques should be particularly useful for analyzing preprocessing for pattern recognition. The selection of measurements becomes, we have found, more meaningful when it is possible to examine relatively homogeneous subsets, i.e., after clustering.

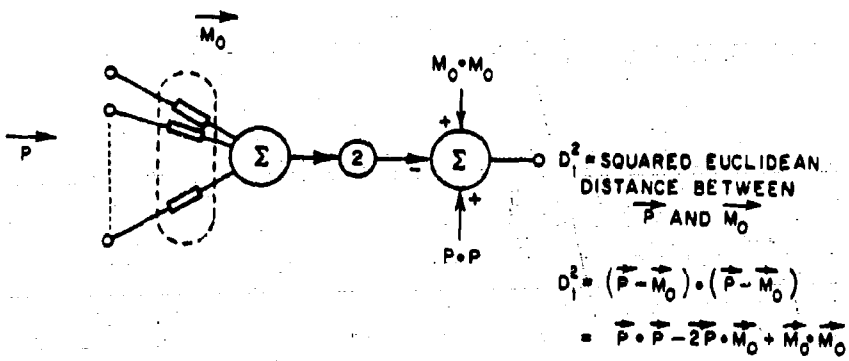
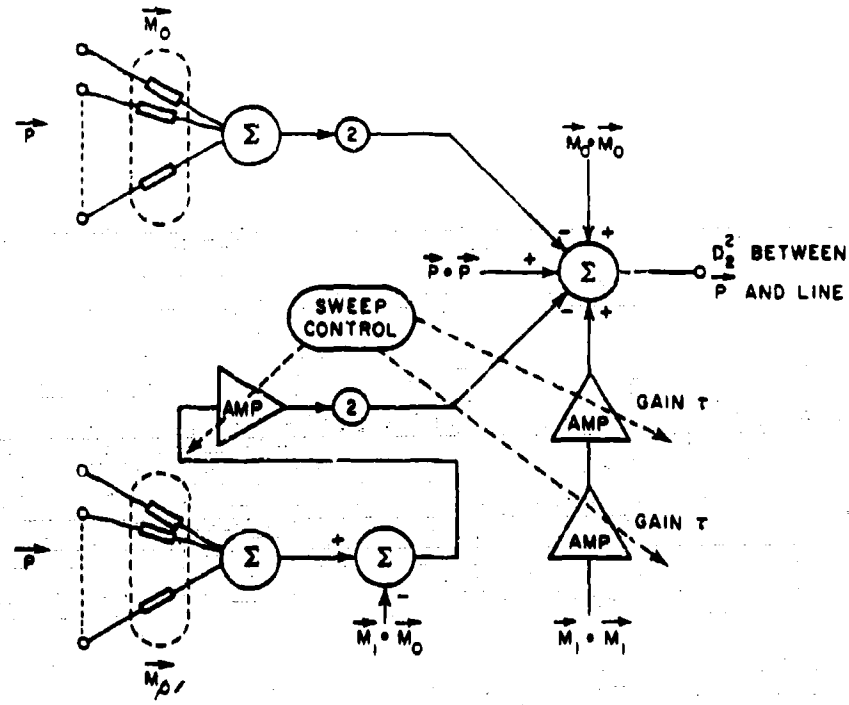


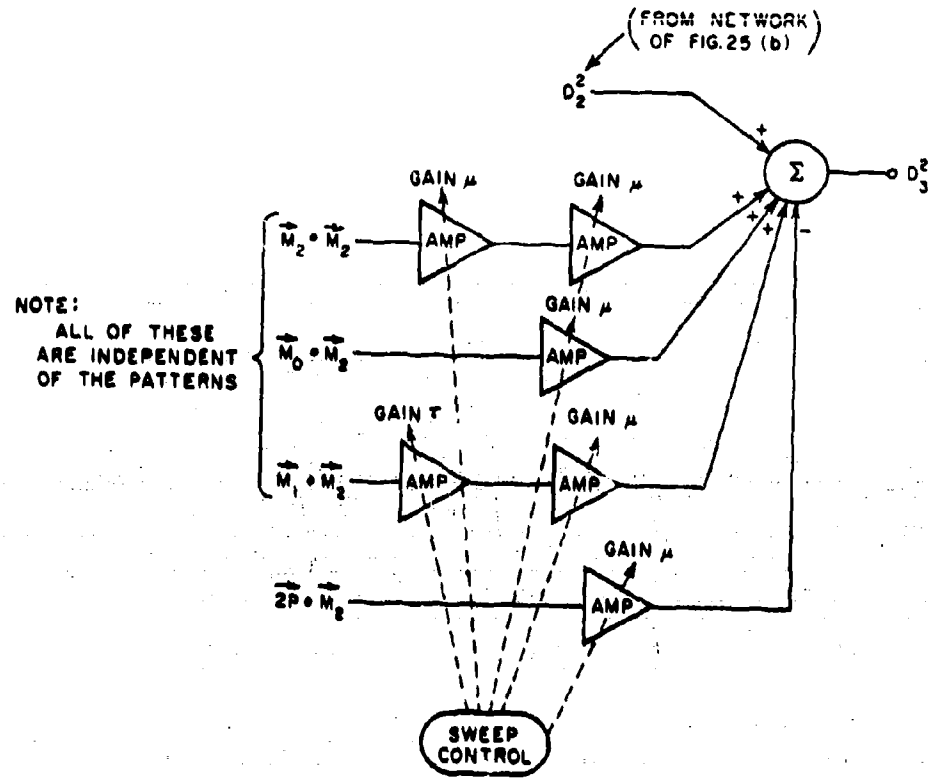
FIG. 25(a) IMPLEMENTATION FOR FINDING EUCLIDEAN DISTANCE OF A PATTERN FROM A POINT



$$\left(\begin{array}{c} D_2^2 \text{ FROM } \vec{P} \\ \text{TO POINT} \\ \text{ON LINE} \end{array}\right) = \left[\vec{P} - (\vec{M}_0 + \vec{M}_1 \tau)\right] \cdot \left[\vec{P} - (\vec{M}_0 + \vec{M}_1 \tau)\right]$$

$$= \vec{P} \cdot \vec{P} - 2\vec{P} \cdot (\vec{M}_0 + \vec{M}_1 \tau) + \vec{M}_0 \cdot \vec{M}_0 + \vec{M}_1 \cdot \vec{M}_1 \tau^2 + 2\vec{M}_1 \cdot \vec{M}_0 \tau$$

FIG. 25(b) IMPLEMENTATION FOR FINDING EUCLIDEAN DISTANCE OF A PATTERN FROM A LINE



$$D_3^2 = [\vec{P} - (\vec{M}_0 + \vec{M}_1 \tau + \vec{M}_2 \mu)] \circ [\vec{P} - (\vec{M}_0 + \vec{M}_1 \tau + \vec{M}_2 \mu)]$$

$$= \vec{P} \circ \vec{P} - 2\vec{P} \circ (\vec{M}_0 + \vec{M}_1 \tau + \vec{M}_2 \mu) + \vec{M}_0 \circ \vec{M}_0 + \vec{M}_1 \circ \vec{M}_1 \tau^2$$

$$+ \vec{M}_2 \circ \vec{M}_2 \mu^2 + 2\vec{M}_0 \circ \vec{M}_1 \tau + 2\vec{M}_0 \circ \vec{M}_2 \mu + 2\vec{M}_1 \circ \vec{M}_2 \tau \mu$$

FIG. 25(c) IMPLEMENTATION FOR FINDING EUCLIDEAN DISTANCE OF A PATTERN FROM A PLANE

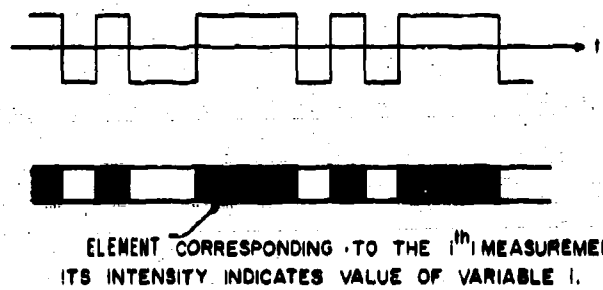


FIG. 26 AN OPTICAL PANEL FOR INPUTTING A HIGH-DIMENSIONAL PATTERN INTO AN ISODATA SYSTEM

We are attempting to improve and refine our techniques for measurement selection and to look at the possibilities of generating meaningful measurements automatically.

VIII ACKNOWLEDGMENTS

We would like to thank the many people at SRI with whom we've worked who have listened to and commented on our ideas. We would particularly like to thank Dean Babcock and Charles Rosen for encouragement and the time to pursue the research necessary to develop these ideas. Discussions with Charles Dawson, Wade Foy, and Nils Nilsson have pointed out important relationships.

The work was supported by internal funding by Stanford Research Institute and contracts from the Graphical Data Transducer Branch, Data Division, Communications Department, USAEL, Fort Monmouth, N.J.

APPENDIX A

ISODATA-LINES AND ISODATA-PLANES

In this appendix we consider the generalization of ISODATA-POINTS to the fitting of connected line segments to "tubular" high dimensional data and to the fitting of triangular segments of hyper-planes to "surface-like" high-dimensional data.*

The fundamental concept of ISODATA is the iterative adjustment of the position of "clusters" in order that these clusters come to reflect, in their relative positions, the structure of the data. In ISODATA-POINTS the goal is to adjust these clusters so that they lie around well-chosen average points.

In the generalizations of ISODATA-POINTS in this section we will relate first pairs (ISODATA-LINES) and then triples (ISODATA-PLANES) of points to each other. In the first generalization we relate pairs of points together in order to create line segments. By allowing each point to be in more than one pair we are able to create a piecewise-linear curve. We propose fitting curves composed of segments of these lines to the data (rather than just single points), thus obtaining an piece-wise linear algebraic expression describing a set of data points in a high-dimensional space. Such iterative fitting of a set of line segments to data we call ISODATA-LINES. Such a curve is shown fitted to a set of hypothetical data in Fig. A-1.

In the second generalization of ISODATA-POINTS we associate triples of points together. These triples of points can be used to define a triangular segment of a plane in n-dimensional space. By allowing points to be in more than one triple of points we link these triangular segments of planes together to form a piecewise-planar surface in n-space. This surface would then be iteratively adjusted to cause it to fit a set of data. This technique we call ISODATA-PLANES Such a surface in a three-dimensional space is shown in Fig. A-2.

We have not developed an adequate algorithm for either of these generalizations. We have investigated ISODATA-LINES to a greater extent, the results of which we now describe.

---

* An example of "tubular" data would be a set of time samples of the patterns at the output of a set of band-pass filters into which a word has been spoken. An example of surface-like data is the values of n-predictors that are used to predict a single predictand such as the ceiling height at Washington, D.C.

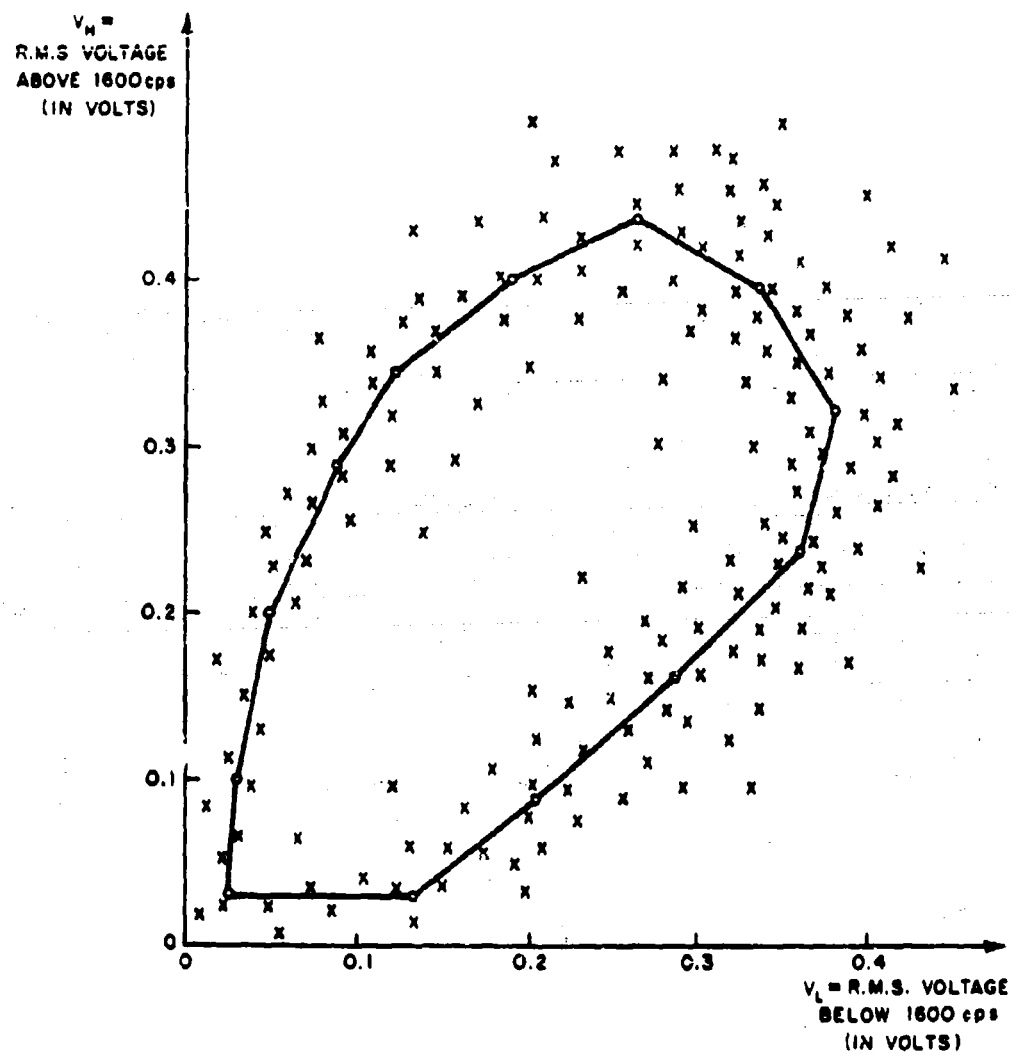


FIG. A-1 AN ISODATA-LINES CURVE FITTED TO A SET OF HYPOTHETICAL DATA
(The Trajectory of the Word "zero")

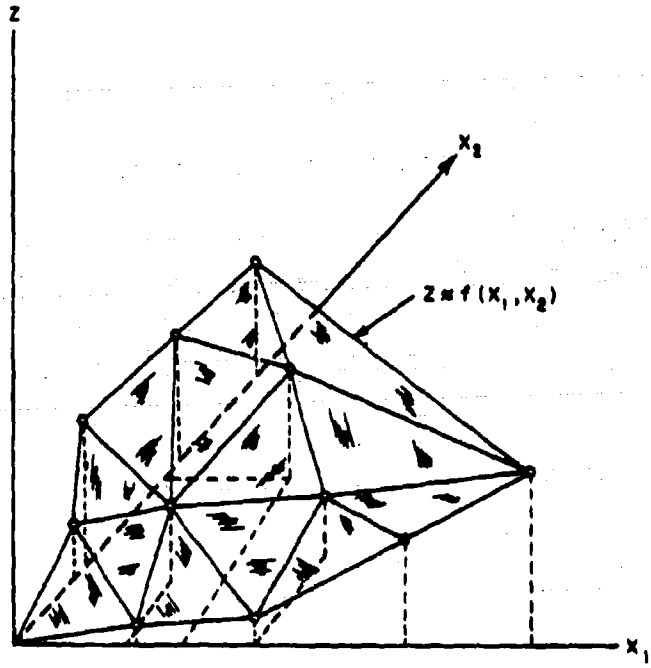


FIG. A-2 AN ISODATA-PLANES SURFACE GIVING Z AS A FUNCTION OF X_1 AND X_2

A. ISODATA-LINES

At the present time it appears to us that the algorithm for ISODATA-LINES should consist of the following steps:

(1) Use ISODATA-POINTS to define cluster centers within the data.

(2) Take the nearest two cluster points and relate them (i.e., they are to define a line segment).*

(3) Starting with one cluster point of this pair, find that new cluster point nearest to it. Relate this pair. (A maximum allowable distance for pairs might be used here.)

(4) Continue this procedure until all acceptable** cluster points are paired. At this point iterative adjustment of the line segments would begin.

(5) Effective iterative adjustment of the line segments requires the answering of the following two questions:

(a) What subset of patterns should be associated (probably not disjointly) with each line segment?

(b) In what direction should each pattern move the cluster points that define the line segment associated with that pattern and what amount should it move it?

---

*Each pair considered for a relationship should be examined to ensure that there are points lying near the straight line segment connecting them. A simple modification of the ISODATA-POINTS program to store the second (and third?) cluster points nearest to the patterns would allow the presence or absence of patterns between two cluster points to be found.

**Some cluster points may be isolated from others due to the nature of the data and in this sense "unacceptable." If this isolated point were split into two points in the manner of ISODATA-POINTS a best-fit line to this isolated cluster could be obtained.

These are two reasonably well-defined questions. Though we have no tested definitive answers, we have the following conjectural answers:

To 5(a); A pattern should be associated with the two line segments to which it is closest.

To 5(b); A pattern should move the cluster points in a direction toward the pattern in a direction perpendicular to the line connecting the extreme end points of two line segments sharing a common center cluster point. In Fig. A-3 this line is shown as a dashed line connecting Cluster Point 1 and Cluster Point 3.

The details of the proposed adjustment procedure (for two dimensions) are shown in Fig. A-4. A perpendicular $\overline{BD}$ is, in effect, erected to line $\overline{AC}$. Patterns associated with these two line segments whose projection (on the line $\overline{AC}$) are in the interval $\overline{AD}$ are used to modify Points A and B. Patterns in the interval $\overline{DC}$ are used to modify Points B and C. A pattern projecting directly on A would modify only A, and the same for C. A pattern projecting on D would modify only B. The proportion of modification made to each of the two points for cases in between would be linear, as indicated in Fig. A-4(b). In Fig. A-4(c) we show the amount of modification made for a sample pattern projected onto $\overline{AD}$.*

Note that the bias in this adjustment procedure tends to straighten the kinks in the piecewise-linear curve shown in Fig. A-3.

As we develop a better understanding of ISODATA-LINES, it seems nearly certain that something analogous to splitting and lumping will be useful. In Fig. A-5 we indicate two situations in which different kinds of splitting might be called for. As for lumping, it could conceivably occur if two cluster points draw close together (i.e., as the line segment between them shortens) or as a cluster point at the end of a curve draws near to a line segment. (Some provision for branching data would be necessary.) Straightness of consecutive segments could also be used as a criterion for dropping the interior cluster point (this is the reverse of the splitting in Fig. A-5(b)).

We feel that the use of the ISODATA-POINTS program to find reasonable starting lines will greatly reduce running time by reducing the number of iterations required by ISODATA-LINES to find a good fit to the data.

---

* The calculation of distance from a line segment does not require more than the algebraic manipulation of the distances from the two defining end points. No explicit formula for the line segment itself need be found.

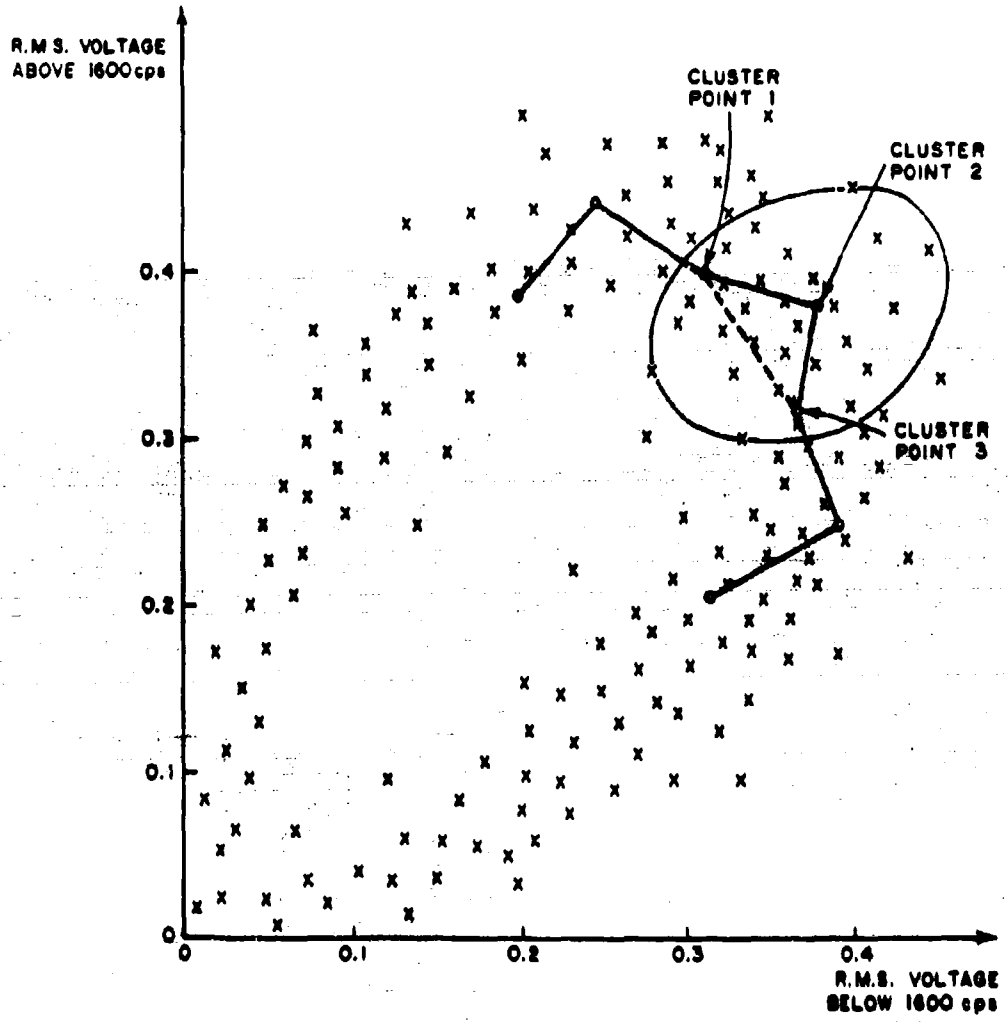
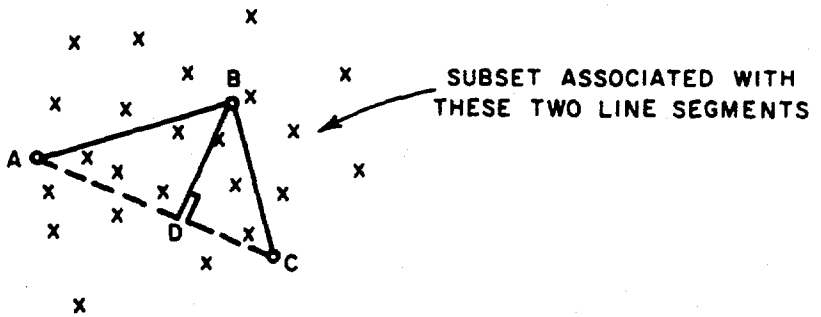
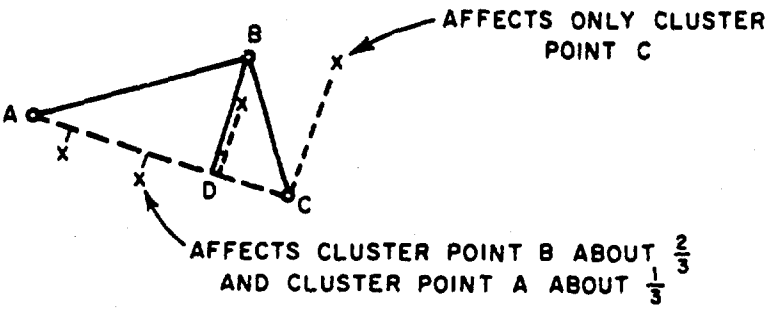


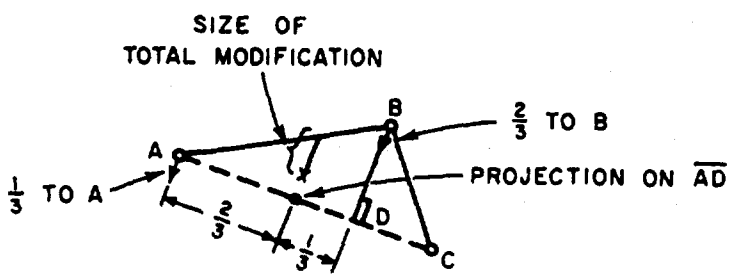
FIG. A-3 CONSIDERATIONS AFFECTING THE ADJUSTMENT OF ISODATA-LINES CURVES



(a)

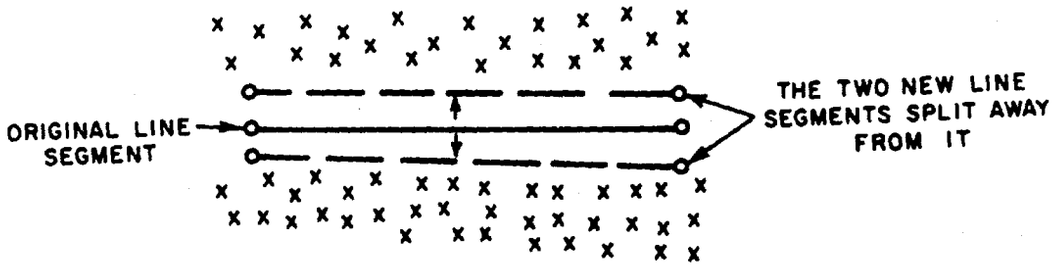


(b)

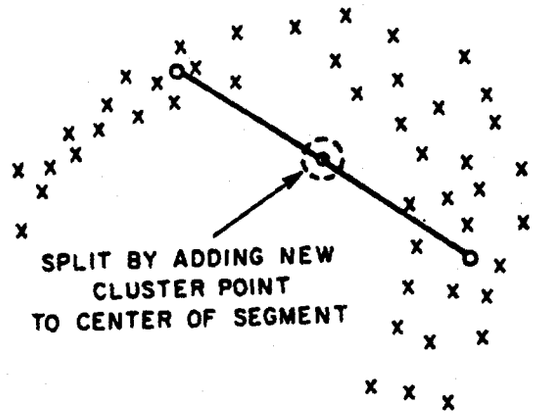


(c)

FIG. A-4(a) THE SUBSET OF PATTERNS ASSOCIATED WITH TWO LINE SEGMENTS
(b) THE CLUSTER POINTS AFFECTED BY PARTICULAR PATTERNS
(c) THE PROPORTION OF CLUSTER POINT MODIFICATION CAUSED BY A SINGLE PATTERN



(a)



(b)

FIG. A-5(a) SPLITTING BY CREATING NEW LINE SEGMENTS
(b) SPLITTING BY CREATING A NEW CLUSTER POINT

B. ISODATA-PLANES

We have not yet worked out a tentative algorithm for the iterative adjustment of these planes, except insofar as the algorithm is similar to ISODATA-LINES. We feel that by the time we have developed ISODATA-LINES we will have a good start on developing ISODATA-PLANES. It is interesting to note that for ISODATA-PLANES we only need distances from the three defining points in order to find the distances from a plane. We do not need an explicit formula for the plane.

REFERENCES

1. Tukey, John W., "The Future of Data Analysis," Annals of Mathematical Statistics Vol. 33, No. 2, p.2, March, 1962. (This is an excellent paper for those concerned with the data analysis aspects of pattern recognition.)

2. Ibid., pp 5-6

3. D.N. Lawley and A.E. Maxwell, Factor Analysis as a Statistical Method, Butterworths, London, p.2, 1963.

4. Samuel S. Wilks, Mathematical Statistics, John Wiley and Sons, Inc., New York, p. 565, 1962.

5. Ibid, p. 544.

6. Bonner, R.E. "On Some Clustering Techniques," IBM J. of Res. and Dev.,p 22-32, Jan., 1964.

7. Cooper, David B., and Cooper Paul W., "Adaptive Pattern Recognition without Supervision," Proceedings of the 1964 IEEE International Convention, 1964.

8. Fierschein, G., and Fischler, M., "Automatic Subclass Determination for Pattern Recognition Applications," Corres. Trans. PGEC, Vol. EC-12, No.2, pp 137-141, April, 1963.

9. Glaser, E.M "Signal Detection by Adaptive Filters," IRE Trans. on Info. Thy., Vol. IT-7, No. 2 pp 87-98 April, 1961.

10. Jakowotz, C.V., Shuey, R.L., and White, G.M., "Adaptive Waveform Recognition," Information Theory, C. Cherry, editor, Butterworths, Washington, D.C., 1961.

11. Rogers, D.J., and Tanimoto, T.T., "A Computer Program for Classifying Plants," Science, 132, pp 1115-1118, 21 Oct. 1960.

12. Sebestyen, George, "Pattern Reccgnition by an Adaptive Process of Sample Set Construction," Trans. PGIT, Vol 8, pp S82-S91, September 1962.

13. Smith, James W., IEEE Trans. on Info. Thy., Vol IT-10, No. 3, pp 208-214, July, 1964.

14. Spilker, J.J., Luby, D.D., Lawhorn, R.D., "Progress Report-- Adaptive Binary Waveform Detection," Communication Sciences Department Technical Report #75, Philco Western Development Lab , Palo Alto, California, December 1963.

15. Stark, L., Okajima, M., Whipple, G.H., "Computer Pattern Recognition Techniques: Electrocardiographic Diagnosis," Communications of the ACM, Vol 5, pp. 527-531, October 1962.

REFERENCES (Continued)

16. Tukey, J.W., op. cit., p.9.

17. Tukey, J.W., op. cit., p. 10.

18. Yates, F., (Discussion) J. Roy, Statist. Soc., Ser. B., Vol 17, p. 31, 1955.

19. Tukey, J.W., op. cit., p13-14.

STANFORD RESEARCH INSTITUTE

MENLO PARK
CALIFORNIA

Regional Offices and Laboratories

Southern California Laboratories
820 Mission Street
South Pasadena, California 91031

Washington Office
808–17th Street, N.W.
Washington, D.C. 20006

New York Office
270 Park Avenue, Room 1770
New York, New York 10017

Detroit Office
1025 East Maple Road
Birmingham, Michigan 48011

European Office
Pelikanstrasse 37
Zurich 1, Switzerland

Japan Office
Nomura Security Building, 6th Floor
1–1 Nihonbashidori, Chuo-ku
Tokyo, Japan

Retained Representatives

Toronto, Ontario, Canada
Cyril A. Ing
67 Yonge Street, Room 710
Toronto 1, Ontario, Canada

Milan, Italy
Lorenzo Franceschini
Via Macedonio Melloni, 49
Milan, Italy